PC program implementing an alternative to the paired t-test which adjusts for regression to the mean

Charles J. Kowalski[a], Emet D. Schneiderman*[b], Stephen M. Willis[b]

[a]*Department of Biologic and Materials Sciences and The Center for Statistical Consultation and Research, University of Michigan, Ann Arbor, MI 48109, USA*
[b]*Department of Oral and Maxillofacial Surgery and Pharmacology, Baylor College of Dentistry, PO Box 660677, Dallas, TX 75266-0677, USA*

Abstract

In many biomedical research contexts, treatment effects are estimated from studies based on subjects who have been recruited because of high (low) measurements of a response variable, e.g., high blood pressure or low scores on a stress test. In this situation, simple change scores will overestimate the treatment effect; and the use of the paired t-test may find significant change due not to the treatment per se but, rather, due to regression towards the mean. A PC program implementing a procedure for adjusting the observed change for the regression effect in simple pre-test-post-test experiments is described, illustrated, and made available to interested readers. The method is due to Mee and Chua *(Am Stat,* 45 (1991) **39–42),** and may be considered as an alternative to the paired t-test which separates the effect of the treatment from the so-called regression effect.

Keywords: Treatment effects; Regression to the mean; Paired t-test; PC program

1. Introduction

The term regression to the mean (or, simply, the regression effect) is used to describe the phenomenon that a variable which is extreme on its first measurement will tend to be closer to the center of the distribution when measured at a later point in time. This phenomenon is exemplified in patients with initially high blood pressure measurements who, if measured on a second occasion, will tend to have more average values at that time. It can be demonstrated in its most simple form if we assume that the first (Z) and second (X) measurements have a bivariate normal distribution with means μ_Z and μ_X, variances σ^2_Z and σ^2_X and correlation ρ. Then the regression function of X on Z, or the conditional expectation of X given that $Z = z$, is

$$E(X \mid Z = z) = \mu_x + \rho \, \frac{\sigma_x}{\sigma_z} \, (z - \mu_z) \tag{1}$$

If we rewrite (1) in the form

$$E(X \mid Z = z) - \mu_x = \rho \, \frac{\sigma_x}{\sigma_z} \, (z - \mu_z) \tag{2}$$

* Corresponding author.

it is seen that, for a given z, the expected distance of X from μ_x (the left-hand side of Eq. 2) is proportional to the distance between z and μ_z, the constant of proportionality being the regression coefficient

$$\beta_{xz} = \rho \ \frac{\sigma_x}{\sigma_z} \tag{3}$$

Thus we would expect X to be closer to its mean than z was to its, if $|\beta_{XZ}| < 1$, i.e., if $|\rho| < \sigma_Z/\sigma_X$.

An important special case of(l), which we will need in our development, has been called dynamic equilibrium [1] and is characterized by equal means and variances for Z and X, viz.,

$$\mu_Z = \mu_X = \mu$$

$$\sigma^2{}_Z = \sigma^2{}_X \doteq \sigma^2$$

and by a correlation which is somewhat less than perfect ($|\rho|$ < 1). This is a 'no change' model which could be used, e.g., in the situation in which a treatment was applied following a baseline measurement and the treatment was completely ineffective. Nevertheless, the regression phenomenon must still be contended with: in this case

$$E(X|Z = z) = \mu + \rho(z - \mu) \tag{4}$$

and, as shown clearly by Lord [1], subjects with sufficiently high (low) values of z are virtually guaranteed to decrease (increase) when the second measurement is taken. Writing this in the form (2)

$$E(X \mid Z = z) - \mu = \rho(z - \mu) \tag{5}$$

it is clear that we expect X to be closer to the mean than z, provided only that $|\rho| \neq 1$. Notice that, if $z > \mu$ and ρ is positive, we still expect that $X - \mu$ will be positive, but at a reduced level.

In many biomedical research contexts the inclusion criteria for a study will necessarily include that subjects have sufficiently high (low) values of the variable of interest, e.g., high values of blood pressure, cholesterol level, or gingivitis; or low scores on an achievement or physical stress test.

Since such subjects can be expected to change even in the absence of treatment, it is important to correct the observed change for the regression effect to properly assess the portion of the change which may be attributed to the treatment. In this paper we describe, illustrate and make available a PC program implementing a procedure which explicitly adjusts for that portion of the observed change attributable to regression to the mean. The procedure was developed by Mee and Chua [2], and we will follow them in describing a scenario in which its use would be appropriate.

2. A typical application

Mee and Chua [2] considered data taken from [3, p. 476], reproduced in Table 1 below, which was gathered in the following context. High school students in Florida are required to pass a literacy test before they can graduate. Students who do not pass the examination on their first attempt may take a refresher course, and then retake an equivalent test. The problem is to decide whether or not the refresher course was effective in raising the test scores.

Notice **first** that it would not be appropriate in this situation to do a simple paired r-test. This test fails to account for change due to the regression effect and thus produces a biased estimate of the effect of the course. As we have seen, even if nothing is done between the **first** and second exams, those individuals who have low scores on the **first** exam will, on the average, show improvement on the second.

Table 1
Literacy test scores before and after taking a remedial course

Student	Before	After	Difference
1	45	49	4
2	52	50	-2
3	63	70	7
4	68	71	3
5	57	53	- 4
6	55	61	6
7	60	62	2
8	59	67	8

Example data. Source: **McClave** and **Dietrich** [3].

We let Z denote a score on the **first** test, and X the corresponding score on the second. Z will be available for all the students in the school; X will be observed only for those students with values of $Z < k$, where k is a suitable cut-off point. Nevertheless, we define μ and σ^2 to be the common mean and variance of the pre- and post-test scores which refer to the situation in which all the students take both examinations. Similarly, ρ is the correlation between the pre- and post-scores when all students take both exams. We let μ_Z and μ_X be the means for the students who actually take both exams; σ^2_Z and σ^2_X are the corresponding variances. Assuming that the refresher course is completely ineffective in raising test scores, because of the regression effect, we do not expect that the differences X − Z will be zero. In fact, the expected difference between the means is

$$\mu_X - \mu_Z = \mu + \rho(\mu_Z - \mu) - \mu_Z \qquad (6)$$

as was noted in [2]. This will be zero only if $\rho = 1$. The property of conditional expectations referred to in [2], necessary to show $\mu_X = \mu + \rho(\mu_Z - \mu)$, and hence (6), is that $\mu_X = \text{E}(X \mid Z < k) = \text{E}[\text{E}(X|Z)|Z < k] = \text{E}[\mu + \rho(Z - \mu)|Z < k] = \mu + \rho(\mu_Z - \mu)$. We see, then, that even in the **absence** of any intervention (the school is in dynamic equilibrium), students selected using the criterion $Z < k$ may be expected to change (improve). The amount of change expected is given by (6). It is this amount of change which must be subtracted from the observed change to obtain the change attributable to the intervention. The method of accomplishing this developed in [2] is outlined next.

3. Mee and Chua's procedure

Mee and Chua consider their method to be an alternative to the paired t-test which more **properly** accounts for the regression effect described above. It is based on regressing X on Z − μ, taking μ, the common value of the means of Z and X under dynamic equilibrium, as known. This will be a reasonable assumption when, in the context of our example, the school is large for then one can use this large sample to estimate the mean of Z.

The null model may be written, from **(4)**, as

$$H_0: X = \mu + \rho(Z - \mu) + \epsilon \qquad (7)$$

and the alternative is

$$H_A: X = \beta_0 + \rho(Z - \mu) + \epsilon \qquad (8)$$

with $\beta_0 > \mu$. This form of the hypothesis (with $\beta_0 > \mu$) is for situations in which the treatment or intervention is intended to raise the scores of treated subjects. We continue to consider this case in this section. We show the (slight) changes necessary to deal with the opposite scenario ($\beta_0 < \mu$) later. In (6) and (7), ϵ is a random error (residual) with $\text{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. We define $\tilde{Z}_i = Z_i - \mu$ and compute

$$\tilde{Z} = \frac{1}{N} \sum \tilde{Z}_i$$

$$\bar{X} = \frac{1}{N} \sum X_i$$

$$\hat{\beta}_1 = \frac{\sum (\tilde{Z}_i - \tilde{Z})(X_i - \bar{X})}{\sum (\tilde{Z}_i - \tilde{Z})^2}$$

$$\hat{\beta}_0 = \bar{X} - \hat{\beta}_1 \tilde{Z}$$

and

$$s^2 = \frac{1}{N - 2} \sum (X_i - \hat{\beta}_0 - \hat{\beta}_1 \tilde{Z}_i)^2$$

The fitted model $\hat{X}_i = \hat{\beta}_0 + \hat{\beta}_1(Z_i - \mu)$ is the least squares regression **line**, and s^2 is the mean squared error of the regression (an estimate of σ^2).

Then the test of H_0 vs. H_A is based on

$$t = \frac{\hat{\beta}_0 - \mu}{s \sqrt{\frac{1}{N} + \frac{\tilde{Z}^2}{\sum (\tilde{Z}_i - \tilde{Z})^2}}} \qquad (9)$$

where t has $N - 2$ degrees of freedom. Note that the test may be rephrased as one of $\beta_0 = \mu$ vs. $\beta_0 > \mu$ so that a one-sided P-value is appropriate and is printed in our program. We also compute and print the $(1 - \alpha)$ x 100% one-sided **confidence** interval for $\beta_0 - \mu$. The user selects the value of $1 - \alpha$ (e.g., 0.95) to be used. The **confidence** interval is given by

$$\beta_0 - \mu \geq (\hat{\beta}_0 - \mu) - t_{1-\alpha} (N - 2)s$$

$$\times \sqrt{\frac{1}{N} + \frac{\tilde{Z}^2}{\sum(\tilde{Z}_i - \tilde{Z})^2}} \qquad (10)$$

where $t_{1-\alpha} (N - 2)$ is the $(1 - \alpha)$th percentile of the t distribution with $N - 2$ degrees of freedom. The right-hand side (RHS) of the inequality in (10) represents a lower bound for the difference $\beta_0 - \mu$. When trying to raise the value of the measurement, we expect $\beta_0 - \mu \geq 0$ and the RHS is then the smallest increase consistent with the data. If this is positive, we may conclude, with confidence $1 - \alpha$, that the TX (treatment or intervention) effect has in fact raised the mean score, and this increase is at least RHS. If this is negative, this indicates that a negative increase (a decrease) is consistent with the data, the interval contains zero, and the TX effect is non-significant.

For the example data in Table 1, the fitted regression function is

$$\hat{X} = 79.959 + 1.1112(Z - \mu)$$

with $s^2 = 20.296$. Mee and Chua took $\mu = 75$ as the true value of μ and showed that this leads to $t = 1.08$, and the corresponding (one-sided) *P*-value is 0.16. The simple paired t-test, on the other hand, gives $t = 2.0$ with $P = 0.04$, which will be judged by most as indicating a significant effect. Thus while the 'total change' may be significant, when the amount of change due to the regression effect is subtracted — leaving that portion of the change which can be attributed to the intervention — the adjusted change need not be. The one-sided 95% confidence interval for $\beta_0 - \mu$ has the lower bound -3.9450. It answers the question, 'How bad

might the treatment be?' In our example, we are 95% confident that the refresher course will not lower the postmeasure by more than 3.945 points. The fact that this change can be negative is consistent with the result of the test based on (9) which showed no significant treatment effect.

4. The **program**

The program is invoked by issuing the command 'gsruni meechua'. The user is **first** asked for the location and name of the (ASCII or GAUSS) file containing the values of the Z/X **(pre/post)** measurements. The data set should be rectangular with N rows and 2 columns; the premeasures in column 1, followed by the postmeasures. The user is then asked to enter the value of μ (MU), and the level of confidence (e.g., 0.95) to be employed in constructing the confidence interval for $\beta_0 - \mu$. Finally, the user indicates whether the TX is expected to INCREASE or DECREASE the mean score for treated subjects.

The output includes plots of the 'growth profiles' **(pre-** and post-measurements connected by straight lines) for each of the N individuals; descriptive statistics (means, standard deviations and correlation coefficients) for Z, X, $D = X - Z$ and Z $W = \tilde{Z}$; *the* regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$; the value of R^2 and the mean square error for the regression of X on Z; the P-value for the test of a significant treatment effect; and the confidence interval for $\beta_0 - \mu$. For the example data, taking $\mu = 75$, the numerical form of the output is as follows:

DESCRIPTIVE STATISTICS

ZBAR = 57.375, SDZ = 6.9885, VZ = 48.8393
XBAR = 60.375, SDX = 8.8146, VX = 77.6964
DBAR = 3.000, SDD = 4.2426, VD = 18.000
ZWBAR = -17.625, SDZW = 6.989,
   VZW = 48.839
R(ZX) = 0.8810
R(ZD) = 0.1831
R(XZW) = 0.8810
RSQ(XZW) = 0.776

In the above, ZW denotes $\tilde{Z}$. The value of

RSQ(XZW) (the square of R(XZW)) measures the strength of the linear relationship between ZW and X, so is useful in judging the goodness-of-fit of the assumed model. At this point in the program we also produce a scatterplot of X vs. ZW to help facilitate this judgement. The descriptive statistics and plot are followed by

ESTIMATES OF PARAMETERS

BETA0 = 79.9590
BETA1 = 1.1112
MSE = 20.2960
SE = 4.5803

MSE is the mean squared error of the regression of X on ZW (s^2). SE is the standard error of $\hat{\beta}_0 - \mu$ as defined by the denominator of expression (9). Following this we plot the fitted regression of X on ZX superimposed on the previous scatterplot and then print

TEST AND CONFIDENCE INTERVAL

OBSERVED VALUE OF t = 1.0827
CORRESPONDING ONE-SIDED P-VALUE
   = 0.1603
THE LOWER BOUND OF THE 95% ONE-SIDED CONFIDENCE
INTERVAL FOR BETA0 − MU IS = -3.9450.
YOU ARE 95% CONFIDENT THAT BETA0
   − MU IS GREATER THAN -3.9450.

Hardware requirements and information concerning the availability of this program, along with other programs for longitudinal data analysis, are given in the Appendix.

5. Discussion

The program described above can be used to adjust or correct simple change scores for the regression effect, and provides a valid test of the null hypothesis that the treatment has had no effect. In order to see more clearly the relationship between the test and the regression effect, it may be helpful to note that H_0: X = $\mu + \rho(Z - \mu) + \epsilon$ implies that E(X) = $\mu_X = \mu + \rho(\mu_Z - \mu)$ which, subtracting μ_Z from both sides, is (6), which is the expected difference between μ_X and μ_Z when only the regression effect is present (dynamic equilibrium). Similarly, H_A: X = $\beta_0 + \rho(Z - \mu) + \epsilon$ with $\beta_0 > \mu$ implies E(X) = $\mu_X = \beta_0 + \rho(Z - \mu)$ which in turn implies that $\mu_X - \mu_Z > \mu + \rho(\mu_Z - \mu) - \mu_Z$, i.e., that the observed change exceeds the regression effect. Thus, H_0 and H_A, previously stated in terms of regression equations, have direct interpretations in terms of the regression effect: H_0 implies that the expected change is due solely to the regression effect; H_A that the expected change exceeds this amount.

Note also that H_A as specified in (8) models an additive treatment effect, i.e., the effect is $\beta_0 - \mu$ no matter what the initial value, Z. This is clear from comparison of (7) and (8): the difference in the expected responses under H_A and H_0 is $\beta_0 - \mu$ since the ρZs cancel out. In certain situations, multiplicative effects, where the effect of the treatment is allowed to depend on the value of Z (so that, e.g., those with the lower scores will show the most improvement) may be more appropriate. Methods which can be used in these situations are described in [4,5]. For a good discussion of the differences between these two types of effects, see [6, p. 68]. Further comparison of (7) and (8), and the method used to obtain the regression of X on $\tilde{Z}$, prompts another remark which may provide some insight into the structure of the technique. $\hat{\beta}_1$ estimates β_1, the population regression coefficient of X on $\tilde{Z}$. This may or may not equal ρ. It will only if $\sigma_Z = \sigma_X$ (cf (3)), and this may not be true if the treatment has had an effect. Mee and Chua [2] suggest that their test will be most sensitive to a treatment effect when $\beta_1 > \rho$; it may be less effective if $\beta_1 < \rho$. In our example, $\hat{\beta}_1 = 1.111$ necessarily exceeds ρ so the test may be presumed to be 'sensitive.'

The reader will have noted that our discussion and example have focused on the situation in which low scores were selected and the treatment was intended to increase the value of the response variable. Our program works equally well when the opposite scenario obtains, i.e., when high values (e.g., blood pressure) are selected and the aim of the intervention is to decrease the value of the measurement. In this case, the user will in-

dicate that the TX is intended to DECREASE the scores of treated subjects. The one-sided **confidence** interval in this case is given by

$$\beta_0 - \mu \leq (\hat{\beta}_0 - \mu) + t_{1-\alpha}(N-2)s$$

$$\times \sqrt{\frac{1}{N} + \frac{\tilde{Z}^2}{\sum(\tilde{Z}_i - \tilde{Z})^2}} \qquad (11)$$

and the RHS is an upper bound for the expected decrease in the mean scores for the selected individuals. Should this be positive, the data do not rule out an increase, and the TX will be judged to be non-significant. If negative, the TX has had a significant effect and the RHS represents the smallest decrease **compatable** with the data, i.e., the reduction will be no less than the RHS.

Finally, we might remark that while the regression phenomenon has been recognized for a long time, it continues to be overlooked in many situations like the one considered in this paper. For a number of additional examples in published research, see [7]. It is hoped that the present discussion, and the availability of software to carry out the associated computations, will have a positive impact on this situation. On the other hand, it needs to be recognized that purely statistical adjustments can never represent a completely adequate substitute for the concurrent study of a control group [8, p. 193]. Such adjustments, including the one described in this paper, should be used only when the use of a control group is not feasible.

Acknowledgement

Appendix

A set of PC programs performing this and related procedures can be obtained on high density 5.25" or 3.5" diskettes (please request type) by sending $25 to defray the cost of handling and license fees. These programs require an **80386-** or **80486-based** personal computer (PC) running the MS-DOS operating system (version 5.0 or higher is recommended, although versions as low as 3.3 will suffice). 80386 computers must also be equipped with an 80387 math coprocessor. At least 4 MB of memory are required, and must be available to **GAUSS386i**, i.e., not in use by memory resident programs such as Windows. EGA or VGA graphic capabilities are required to display the color graphics; VGA or SVGA is suggested to display optimally the graphic results. **Runtime** modules are supplied with the programs so that no additional software (i.e., compiler or interpreter) is required to run these programs. One can create and edit ASCII data sets for use by these programs using the full screen editor supplied with MS-DOS version 5.0. The programs are written and compiled using **GAUSS386i**, version 3.0, require no additional installation or modification, and are run with a single command. When requesting the programs, address inquiries to the corresponding author and make checks payable to Baylor College of Dentistry.

References

[1]   Lord FM: Elementary models for measuring change. In: *Problems in Measuring Change* (Ed: CW Harris), Univ. Wisconsin Press, Madison, 1963.

[2]   Mee RW and Chua TC: Regression toward the mean and the paired sample t test, *Am Stat, 45* (1991) 39-42.

[3]   McClave JT and Dietrich FH: *Statistics,* 4th ed., **Dellen,** San Francisco, 1988.

[4]   James KE: Regression toward the mean in uncontrolled clinical studies, *Biometrics, 29 (1973)* 121-130.

[5]   Senn SJ and Brown RA: Estimating treatment effects in clinical trials subject to regression to the mean, *Biometrics, 41 (1985) 555-560.*

[6]   Hills M: *Statistics for Comparative Studies,* Chapman and Hall, London, 1974.

[7]   Nesselroade J, Stigler S and Baltes P: Regression towards the mean and the study of change, *Psychol Bull, 88 (1980) 622-637.*

[8]   Fleiss JL: *The Design and Analysis of Clinical Experiments,* Wiley, New York, 1986.