

Regression to the mean in pretest-posttest comparisons with two-level selection of subjects

Taichi Okumura*

(平成24年 9 月27日受付；平成24年10月18日受理)

Abstract

This article discusses regression to the mean (RTM) artifacts in the two-level pretest-posttest designs with no control groups when some data are not available for selection based on the pretest scores. A simulation study shows that ignoring missing data will cause the improvement in scores over time to be evaluated incorrectly under several different specifications. Researchers should take RTM caused by data selection into account if they intend to evaluate treatment effects with high internal validity. Further analytical investigation and the development of statistical techniques for adjusting RTM remain for future work.

KEY WORDS :

regression to the mean, pretest-posttest design, multilevel analysis, incomplete data

1. Introduction

Statisticians argue that regression to the mean (RTM) artifacts are important for analyzing data using a pretest-posttest design (Bonate, 2000). When X and Y are the pretest and posttest scores, the conditional expectation of Y given $X = x$ is

$$E(Y | X = x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x). \quad (1)$$

Hence, it can be shown that

$$|E(Y | X = x) - \mu_y| < |x - \mu_x| \quad (2)$$

if $\sigma_x = \sigma_y$. The posttest scores are therefore expected to be nearer to the mean than the corresponding pretest scores. That is, students who got lower marks will receive higher marks on subsequent tests in expectation even without any actual improvements, and vice versa. This result also applies in cases where $\sigma_x \neq \sigma_y$ if the two variables are standardized and the deviations are evaluated in the standardized unit.

Chuang-Stein and Tong (1997) show that the conditional expectation of the difference score $Y - X$ given $X \geq \mu + c\sigma$ is

$$E(Y - X | X \geq \mu + c\sigma) = -(1 - \rho) \frac{\Phi(c)}{1 - \phi(c)} \sigma \quad (3)$$

when X and Y are distributed as a bivariate normal distribution with equal means and variances. $\Phi(\bullet)$ and $\phi(\bullet)$ are the probability density function and the cumulative distribution function of the standardized normal distribution, respectively. Eq. (3) says that the expected difference score is negative, but it is actually zero if X is restricted to be higher than $\mu + c\sigma$. In other words, a researcher may incorrectly interpret the results to imply that posttest scores improved if he does not consider RTM when subjects are selected based on pretest scores.

RTM has been a critical problem in the medical sciences in particular. Researchers have proposed some techniques for evaluating treatment effects in pretest-posttest designs when subjects are selected based on the pretest scores. For example, Mee and Chua (1991) propose a t -statistic for testing the significance of the mean

*School Education

difference when subjects are selected based on their pretest scores with a known population mean. This technique has been extended to situations where the population mean is unknown (George, Johnson, Shahane, & Nick, 1997; Ostermann, Willich, & Lüdtke, 2008). On the other hand, social scientists, including educational researchers and psychologists, do not seem to have seriously considered RTM as a methodological issue even though RTM itself is usually explained in introductory statistics courses (Campbell & Kenny, 2003).

The remainder of the paper proceeds as follows. In the next section, the author indicates that subjects are often selected in multiple stages in the educational sciences. Then, a simulation study is conducted to investigate RTM under the multilevel selection of subjects. Finally, the author argues the importance of taking RTM caused by multilevel selection into account for evaluating the effectiveness of treatments by using pretest-posttest design without control groups.

2. Selection of subjects in multiple stages

The pretest-posttest design is widely used in the educational sciences for evaluating the effectiveness of various types of instruction. There are several noteworthy characteristics of the data gathered in these fields. First, control conditions are not always adopted. RTM can be cancelled in expectation by creating an equivalent control group in addition to the experimental groups and comparing improvements between the groups, since RTM itself is considered to be constant. However, it is difficult in practice to create equivalent control groups in educational settings, since students belong to existing groups such as schools and classes, and in many cases, researchers can only assign entire groups to experimental treatments. Therefore, the control groups are usually not equivalent to the experimental groups.

Moreover, even these non-equivalent control groups are sometimes excluded from an experiment for ethical reasons; for example, it is unfair to leave some students untreated. Such experiments have a one-group pretest-posttest design (Shadish, Cook, & Campbell, 2002). For example, Sasaki and Sugawara (2009) measure the effectiveness of structured group encounters (SGE) for Japanese primary students using this design. Iwataki (2010) also uses this design to evaluate role-playing and social skills training in moral education. As a matter of fact, a researcher can at least partly avoid ethical problems by giving the experimental treatment to the control groups after the posttest, though educational researchers rarely seem to do so.

Secondly, as noted above, educational data are often gathered in a multilevel fashion. That is, schools are the primary sample unit, and students are sampled from the schools. In other words, each observation is nested within a higher sampling unit. It is well known that such data should be analyzed taking the multilevel structure into account. For example, Firebaugh (1978) indicates that an analysis can find opposite correlations depending on whether such structures are considered or dismissed. Kurita (1996) shows that statistical power declines when a standard t -test is applied to hierarchically structured data through Monte-Carlo simulation studies. As noted in the previous section, many articles have indicated the importance of evaluating RTM in pretest-posttest designs using simple random sampling. However, if subjects are sampled hierarchically, selection can occur at all levels. That is, both negatively evaluated schools and poorly graded students may be sampled from the population. Hence, it is important to investigate how influential RTM can be when it is not properly accounted for and data are hierarchically screened. The following section describes the simulated case where the multilevel selection of subjects is not taken into account in a pretest-posttest design without control groups.

3. Simulation design

3.1. Data generation

Now let X_g and Y_g denote the pre- and posttest observations in the g th group ($g = 1, \dots, G$). These values are generated from the bivariate normal distribution:

$$\begin{pmatrix} X_g \\ Y_g \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_{X_g} \\ \mu_{Y_g} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right). \quad (4)$$

Furthermore, the location parameters μ_{X_g} and μ_{Y_g} are independently and normally distributed as

$$\begin{pmatrix} \mu_{X_g} \\ \mu_{Y_g} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \right). \quad (5)$$

This formulation hypothesizes that the mean difference is zero in the overall population, though group means can take various values.

In this simulation, the parameter values were set to be $\mu=50$ and $\sigma^2 + \tau^2=100$. The variances were divided into two parts, and the interclass correlation coefficient ($ICC: \tau^2/\sigma^2 + \tau^2$) took one of three values ($ICC=0.2, 0.5, 0.8$). The correlation coefficient took one of two values ($\rho=0.2, 0.5$), which reflect the low and moderate case, respectively. This study assumed four screening types; that is, all cases were observed, individuals with $X < c$ were observed, all individuals nested within groups where $\bar{X}_g < c$ were observed, and individuals who satisfied $X < c$ and belonged to the groups with $\bar{X}_g < c$ were observed. Hence, 24 conditions were specified in all, as shown in Table 1. In this study, the threshold c was set to be 60. For each specification, pretest and posttest data were generated for 20 individuals and 50 groups (1,000 cases if completely observed), and 5,000 replications were conducted.

Table 1
Specifications for Data Generation

ICC	ρ	Selection		ICC	ρ	Selection	
		Individual	Group			Individual	Group
0.2	0.2	✓	✓	0.5	0.2	✓	✓
0.2	0.2	✓		0.5	0.2	✓	
0.2	0.2		✓	0.5	0.2		✓
0.2	0.2			0.5	0.2		
0.2	0.5	✓	✓	0.5	0.5	✓	✓
0.2	0.5	✓		0.5	0.5	✓	
0.2	0.5		✓	0.5	0.5		✓
0.2	0.5			0.5	0.5		
0.2	0.8	✓	✓	0.5	0.8	✓	✓
0.2	0.8	✓		0.5	0.8	✓	
0.2	0.8		✓	0.5	0.8		✓
0.2	0.8			0.5	0.8		

3.2. Data analysis

For estimating RTM, the following model was applied to completely observed cases:

$$D_{ig} = \delta_g + r_{ig} \quad (6)$$

$$\delta_g \sim N(\delta, \tau_\delta^2) \quad (7)$$

$$r_{ig} \sim N(0, \sigma_\delta^2) \quad (8)$$

where D_{ig} denotes the i th individual's score change ($Y_{ig} - X_{ig}$) in the g th group. The mean difference in this model corresponds to δ , which has a value of 0 since the means of the pretest and posttest are equivalent overall, as shown in the data generation model.

The parameters δ , σ_δ^2 , and τ_δ^2 were estimated using the restricted maximum likelihood (REML) method (Raudenbush & Bryk, 2002). For each dataset, the null hypothesis concerning δ was tested at the 5% significance level using the Wald statistic introduced in Snijders and Bosker (2012). R-2.15.0 (R Development Core Team, 2012) was used for generating the data and fitting the model to the data. The REML estimation was performed by the `lme` function in the `nlme` package.

4. Results of the simulation study

Figures 1 and 2 show the averaged estimates of δ and the practical significance levels on the Wald tests for δ (i.e., the percent of times that the Wald test incorrectly rejected the null hypothesis).

When the ICC was low ($=0.2$), the estimates of δ did not seem to be seriously biased, if the data were screened only at the group level. The practical significance levels (6.0 to 6.6%) were not so different from the theoretical value. However, RTM became very influential when the data was screened at the individual level and ρ was small. In these cases, more than 70% of the Wald tests erroneously rejected H_0 when $\rho=0.2$.

When the ICC was moderate ($=0.5$), RTM caused fairly biased estimates and incorrect significance levels even if the data were screened only at the group level. The degree of bias did not seem to be related to the level of correlation between the two occasions. If the data were screened at the individual level, the estimation biases caused by RTM increased when ρ declined, as in the case when ICC was small. However, the practical significance level of 6.5% was close to the theoretical value only when the pretest and posttest scores were strongly correlated ($\rho=0.8$).

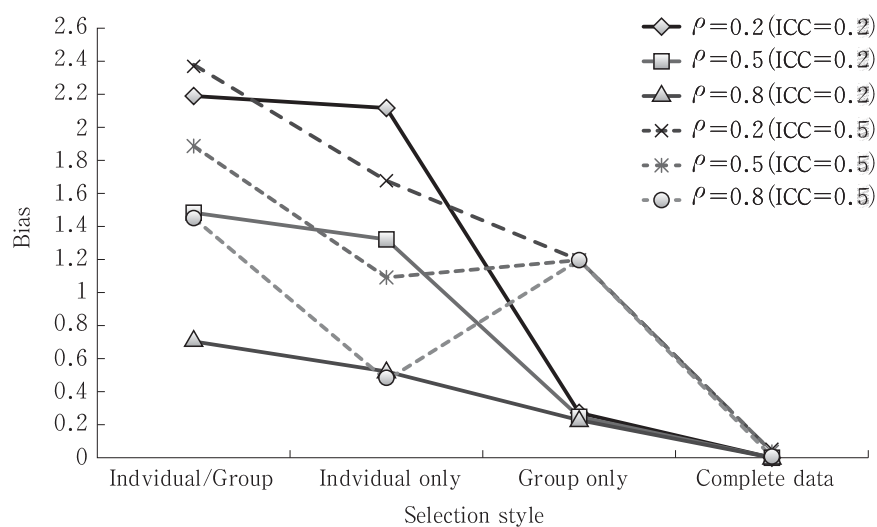


Figure 1. Selection style and bias.

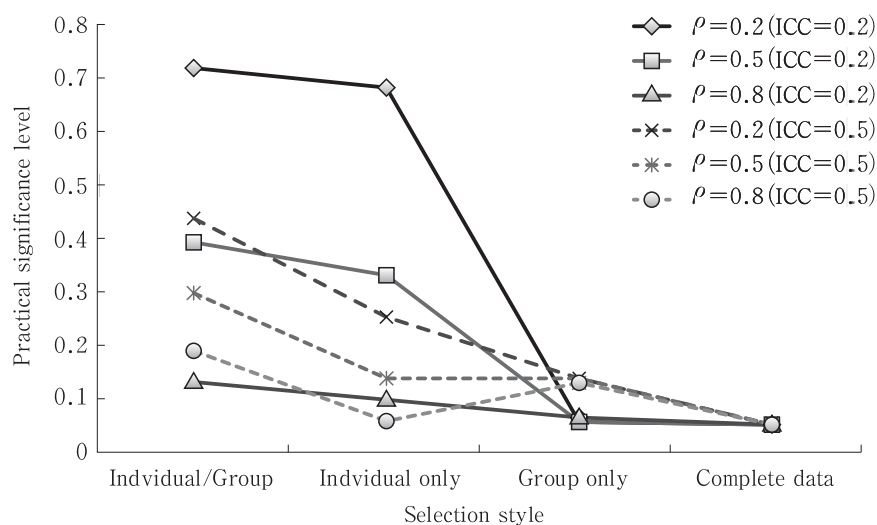


Figure 2. Selection style and significance level.

5. Discussion

The results imply that multilevel data selection based on pretest scores can affect the evaluation of treatment effects under various specifications. Ignoring individual level screening when the ICC is low or group level screening when the ICC is high can seriously inflate the estimation bias and the practical significance level when evaluating treatment effects using incomplete pretest and posttest data.

These results heavily depend on the parameter values chosen to generate the data, as described in section 3.1. For example, under the given specifications, group level screening itself does not produce large RTM artifacts when $ICC=0.2$. However, the practical significance level grows to 90% if the threshold is specified to be 50. In short, if data are available only for individuals or groups whose (mean) pretest scores are lower than the grand mean, it is highly probable that a significance test will incorrectly conclude that the treatment was effective. These results do not guarantee that RTM can be ignored for certain values of the ICC. Instead, RTM may have a critical effect on the results even if the thresholds or other parameters are changed only slightly from the specifications in this paper.

Several tasks remain for future work. Statistical techniques for adjusting RTM when the data have a multilevel structure and data from multiple stages are missing for some subjects should be developed for use in data analysis. In addition, researchers do not always select specific subjects based on such simple cutting criteria. Instead, they include certain students or schools in their samples because these students or schools seem to be suitable for some types of instruction or educational programs or need the treatment to improve. Further research should consider this non-systematic selection of subjects as well.

Acknowledgements

The author presented this study in part at the 40th meeting of Behaviormetric Society of Japan held in Niigata on 16 September 2012. This study was supported by MEXT/JSPS KAKENHI Grant Number 24730530 and Joetsu University of Education Research Grant.

References

- Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman & Hall/CRC.
- Campbell, D. T., & Kenny, D. A. (2003). *A primer on regression artifacts*. New York, NY: The Guilford Press.
- Chuang-Stein, C., & Tong, D. M. (1997). The impact and implication of regression to the mean on the design and analysis of medical investigations. *Statistical Methods in Medical Research*, 6, 115-128.
- Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*, 43, 557-572.
- George, V., Johnson, W. D., Shahane, A., & Nick, T. G. (1997). Testing for treatment effect in the presence of regression toward the mean. *Biometrics*, 53, 49-59.
- Iwataki, D. (2010). Discussion on moral teaching in special needs class: Practice with role playing and social skills training. *Journal of the Tokyo University of Marine Science and Technology*, 6, 13-23. (in Japanese with English abstract)
- Mee, R. W. & Chua, T. C. (1991). Regression toward the mean and the paired sample *t* test. *American Statistician*, 45, 39-42.
- Kurita, K. (1996). The biasing effects of violating the independence assumption upon the power of *t* test. *Japanese Journal of Educational Psychology*, 44, 234-242. (in Japanese with English abstract)
- Ostermann, T., Willich, S. N., & Lüdtke, R. (2008). Regression toward the mean: A detection method for unknown population mean based on the Mee and Chua's algorithm. *BMC Medical Research Methodology*, 8, 52.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- R Development Core Team (2012). *A language and environment for statistical computing*. R Foundation for Statistical

Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Sasaki, M., & Sugawara, M. (2009). The efficacy of structured group encounter in elementary school. *Journal of Clinical Research Center for Child Development and Educational Practice*, 8, 107-117. (in Japanese)

Snijders, T. A. S., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.