Comparing Apples with Apples: Assessing Student Attitudes in the Presence of Regression to the Mean

Anne Michele Millar[1], Bethany JG White[2], Rebecca Romo[3]

[1]Department of Mathematics and Computer Science, Mount Saint Vincent University,
Halifax, Nova Scotia, B3M 2J6
[2]Department of Statistical & Actuarial Sciences, The University of Western Ontario,
London, Ontario, N6A 5B7
[3]Department of Mathematics and Computer Science, Mount Saint Vincent University,
Halifax, Nova Scotia, B3M 2J6

Abstract

When we assess students' attitudes towards statistics we are typically more interested in how their attitudes have changed during the course, than their attitudes at the end of the course. This change, or gain, in attitude scores (post-score - pre-score) depends on the pre-score: negative attitudes tend to have a higher gain score, while more positive attitudes tend to have lower ones. If we wish to compare results across groups of students, we need to consider their gains relative to their pre-scores for a meaningful comparison – apples with apples. Since attitude scores are subject to measurement error, the "true" relationship between the gain and the pre-score is difficult to estimate due to "regression to the mean". In this paper, we compare methods for parameter estimation and confidence interval construction to accurately estimate gains in attitude scores relative to pre-scores in the presence of measurement error.

Key Words: regression to the mean, statistics education, Survey of Attitudes toward Statistics, SATS, students' attitudes, pre-post change

1. Introduction and Motivation

The Student Attitudes Toward Statistics Survey (SATS), Copyright C. Schau (1996, 2003) includes 36 items that assess six components of students' attitudes as well as other items to measure demographic information of the students. Components include Affect, Cognitive Competence, Value, Difficulty, Interest, and Effort. A Likert-type scale that ranges from 1 ("Strongly disagree") to 7 ("Strongly agree") is used for each of the 36 items. A component score is calculated through averaging the scores of all items for that component. All component-specific items must be completed for a student to receive a score on that particular component. Information about all six attitude components and the additional constructs assessed by the SATS is found on the SATS website (Schau, 2005). In this paper we will focus on the Affect component. A student's Affect score reflects his or her "feelings concerning statistics." This is scored as the average of six items (e.g. "I will enjoy taking a statistics course", " I am scared by statistics").

When change in scores is of interest, which is often the case, students' attitudes are assessed with this survey during the first week of the semester to obtain pre-scores. The survey is administered again at the end of the semester to collect information on post-scores. Change (or gain) scores can then be computed as Gain = Post-score – Pre-score. The magnitude of the gain score depends on the pre-score. If the pre-score is low then there is more room for increase (i.e. a higher gain). In contrast, if the pre-score is high, then there is more potential for a drop in score (i.e. a lower gain). This is supported visually using SATS data in Figure 1 where the same data (n=198) are shown in two different formats (Millar and Schau, 2010).

In Figure 1, the left-hand plot displays pre-scores versus post-scores, with the identity (y = x) line. The right hand plot shows pre-scores versus gain scores, with a horizontal (y = 0) line. It clearly shows a negative relationship between the observed pre-scores and gain scores. In each plot, the vertical distance from the line gives the observed gain. Most students with lower pre-scores (e.g. 2) had a higher post-score, while most students with higher pre-scores (e.g. 6) had lower post-scores (see Millar and Schau, 2010). Therefore, in order to draw meaningful conclusions, gain scores should be considered relative to pre-scores, not on their own.

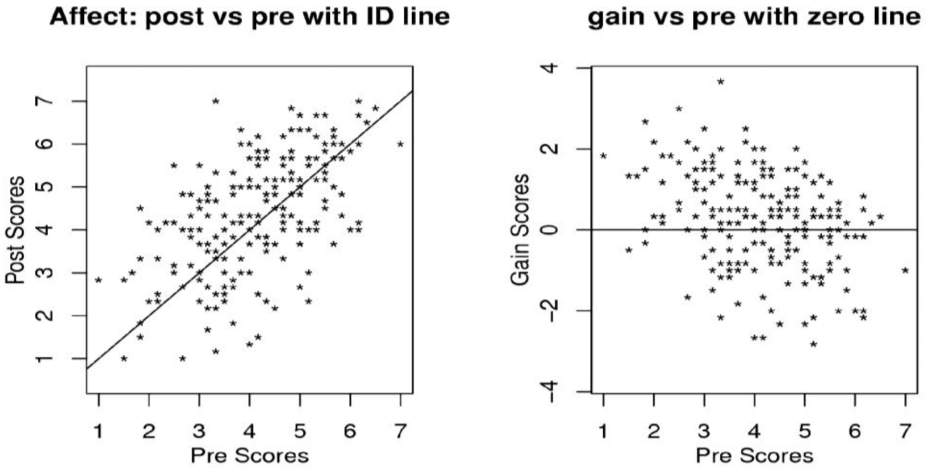

Figure 1: SATS Data Set (n=198) -Post-scores versus Pre-scores, and Gain versus Pre-scores for Affect.

The pre-scores include measurement error, as evidenced by a non-zero test-retest coefficient (or reliability ratio) for the survey, ρ_{xx}, where x is the pre-score. Hence, the negative relationship we observe between the gain scores and the pre-scores may be entirely a consequence of the regression to the mean phenomenon. The purpose of this research is to explore this further to see whether there is an underlying negative relationship between the “true” gain (i.e. the expected gain score for each individual student) and the “true” pre-score, and, if so, if this “true” relationship can be estimated.

2. Synopsis of Research

In this paper, we will use the following notation to distinguish between true and observed scores:

- y_{true} = expected post-score, and $y_{obs} = y_{true} + \delta_y$ = observed post-score,
- x_{true} = expected pre-score, and $x_{obs} = x_{true} + \delta_x$ = observed pre-score,
- $gain_{true} = y_{true} - x_{true}$ = expected gain, and $gain_{obs} = y_{obs} - x_{obs}$ = observed gain

The associated linear regressions are:

$$y_{obs} = \beta_0 + \beta_1 x_{obs} + \varepsilon_{obs} \quad \rightarrow \quad \text{gain}_{obs} = \beta_0 + (\beta_1 - 1)x_{obs} + \varepsilon_{obs} \tag{1}$$

$$y_{true} = \alpha_0 + \alpha_1 x_{true} + \varepsilon_{true} \quad \rightarrow \quad \text{gain}_{true} = \alpha_0 + (\alpha_1 - 1)x_{true} + \varepsilon_{true} \tag{2}$$

For the purposes of this work, we assume the various ε's and δ's are all mutually independent, normally distributed with mean zero, and that δ_y and δ_x have equal variance.

We also proceed under the assumption that there is not a positive relationship between the gain and the pre-score: either the variables are independent or there is a negative relationship. If the slope for the true gain in (2) (i.e. α_1 -1) is negative (or equivalently, the slope for the regression of the expected post-scores on the expected pre-scores, α_1, is less than one), then we have identified a negative relationship. The question then becomes: how can we estimate the values of α_o and α_1, the parameters describing the relationship between true scores, based on observed scores? Here we investigate two methods: a classical errors-in-variables model (EIV); and a model based on an adjusted post-score.

2.1 Methods of Adjustment

2.1.1 Errors-in-variables Model (EIV)

Substituting $x_{obs} = x_{true} + \delta_x$ into (1) gives $y_{obs} = \alpha_0 + \alpha_1 x_{true} + \varepsilon_{true} + \delta_y$ where the error term $\varepsilon_{true} + \delta_y$ is independent of the observed pre-score. This resembles a classical errors-in-variables (EIV) problem, so we can estimate the true slope from the observed slope using the relationship $\alpha_1 = \dfrac{\beta_1}{\rho_{xx}}$ where ρ_{xx} is the test-retest coefficient (Fuller, 1987). Under standard ordinary least squares assumptions, $\hat{\beta}_1$ is an unbiased estimator for β_1 so we can obtain an unbiased estimator for α_1 using this relationship:

$$\hat{\alpha}_1 = \frac{\hat{\beta}_1}{\rho_{xx}} \text{ or } \frac{\hat{\beta}_1}{\hat{\rho}_{xx}} \tag{3}$$

2.1.2 Adjusted Post-score method

The post-scores are adjusted using $y_{adj} = y_{obs} - \dfrac{\sigma_{yobs}}{\sigma_{xobs}}(\rho_{xx} - 1)(x_{obs} - \mu_x)$ and observed pre-scores are then regressed on this adjusted post-score (Roberts, 1980). In practice, the σ's, μ_x and ρ_{xx} are unknown so their estimates are used instead (see (4)).

$$y_{adj} = y_{obs} - \frac{s_y}{s_x}[\hat{\rho}_{xx} - 1](x_{obs} - \bar{x}_{obs}) \tag{4}$$

2.1.3 Estimating the Test-retest Coefficient, ρ_{xx}

Both adjustment methods require an estimate of the test-retest coefficient, ρ_{xx}. The test-retest coefficient (ρ_{xx}) is defined formally as the expected correlation between observed scores (x_1 and x_2) when an individual completes a survey for a first and second time (Fuller 1987).

A reliability study could be conducted to estimate ρ_{xx} for each of the SATS components. Ideally, this would mean that students who are not taking a statistics course would complete the survey twice about a few weeks apart. As this was not practical, we computed estimates based on a large database of SATS scores of students from different institutions, courses and sections. A total of 60 students whose scores are stored in this database completed the pre-survey twice. The test-retest reliability for each SATS component was estimated as the correlation between the two sets of these pre-scores (see Table 1). These represent rough estimates of the test-retest reliabilities but do provide some insight on the extent of measurement error in the SATS components. They are also consistent with reliabilities reported for other attitude scales. For instance, reliability studies have been conducted for Wise's 1985 *Attitudes Toward Statist*ics (ATS) scale based on different student populations and follow-up times. They resulted in reliability estimates ranging between 0.59 to 0.91 for the *Field* subscale and 0.72 to 0.82 for the *Course* subscale (Schultz & Koshino,1998; Vanhoof et al., 2006). A large reliability study on the *Attitudes Toward Mathematics Inventory* (ATMI) estimated reliabilities as between 0.70 and 0.88 for the four subscales based on a 4 month follow-up period (Tapia & Marsh, 2004).

Table 1: Estimated Test-retest Reliabilities
for SATS-36 Components

Component	$\hat{\rho}_{xx}$
Affect	0.79
Cognitive Competence	0.81
Value	0.87
Difficulty	0.88
Interest	0.72
Effort	0.74

In attitude research the test-retest coefficient has been estimated using the correlation between the observed pre and post scores (e.g. Rocconi, & Ethington, 2009). Although we include this in our simulations, we do not recommend this estimate for SATS data. Instead we explore treating this correlation, $\text{cor}(y_{obs},x_{obs})$, as a lower bound and using a half-way, or M-W (Millar-White) estimate:

$$\frac{\text{cor}(y_{obs}, x_{obs})+1}{2} \tag{5}$$

2.2 Simulation Study

To compare the various methods we ran a series of simulations with a number of values for the true slope ($\alpha_1 = 0.7, 0.8, 0.9, 1.0$) and the test-retest coefficient ($\rho_{xx} = 0.7, 0.8, 0.9, 1.0$), generating all data from normal distributions. Sample sizes of n=50, and n=200 were investigated.

For the first series of simulations, we assumed the ratio of the variance of the post scores to the variance of the pre-scores was 1.25 (this value was motivated by a SATS data set from a large university in the U.S.), and that the scores have the same mean. Then, without loss of generality, we assumed $Y_{obs} \sim N(\mu = 4, \sigma^2 = 1.25)$, $X_{obs} \sim N(\mu = 4, \sigma^2 = 1.25)$ for simulation purposes. Based on these values, the remaining relevant parameter values were computed.

Our simulations proceeded as follows:

1. Generate x_{true}, y_{true}, δ_x , δ_y , ε and use these values to determine y_{obs} x_{obs}
2. Calculate estimates and confidence intervals for α_1 based on seven different methods:
    a) Regress the "true" post-score on the true pre-score (benchmark)
    b) Regress the observed post-score on the observed post-score (naïve)
    c) EIV method using the correct value of ρ: $\mathbf{E}\rho$
    d) EIV method using the M-W halfway estimate: $\mathbf{E}\hat{\rho}$
    e) The adjusted method using the correct value of ρ: $\mathbf{A}\rho$
    f) The adjusted method using the correlation between observed pre-and post-scores: **A**cor
    g) The adjusted method using the M-W halfway estimate: $\mathbf{A}\hat{\rho}$

We did not adjust standard errors to construct confidence intervals. For methods c) and d), the naïve margin of error (method b) was used. Therefore, the margin of error was the same for all methods. Each set of simulations was based on 10,000 trials.

Table 2: Simulation Results based on $\alpha_1 = 0.7$, $\rho_{xx} = 0.7$, and n=50

	True	*Obs*	$\mathbf{E}\rho$	$\mathbf{E}\hat{\rho}$	$\mathbf{A}\rho$	Acor	$\mathbf{A}\hat{\rho}$
MSE	0.018	0.063	0.041	0.025	0.044	0.196	0.026
Coverage Level (nominal 99%)	0.995	0.882	0.937	0.973	0.941	0.383	0.994

Based on the simulation results in Table 2, we see that the mean square errors (MSE's) for methods using the M-W half estimates are superior to those using the correct values for ρ_{xx} (a somewhat surprising result), and that the correlation was not a good substitute for the test-retest coefficient. However, these results were not consistent across parameter configurations. Depending on the values of true slope and the test-retest coefficient, and the sample size, the performance rankings of the methods change, sometimes quite dramatically. We will discuss this in further detail in a future publication.

2. Application and Conclusions

Since there does not seem to be a consistent best choice of methods when dealing with measurement error in SATS data, we instead suggest a simulation-based decision of the appropriate methods based on features of the given data set. When analyzing a SATS data set, our approach is as follows:

1. Estimate ρ_{xx}, and then estimate α_1 using each method to obtain initial estimates.
2. Run the simulation code (which will be available on the first author's website in early 2014) with these values, along with the ratio of the observed variances, and the sample size.
3. Identify the "best" method based on MSE and coverage levels.

For the affect component of our SATS data we found the "best" method differed depending on the initial estimate of α_1. In Table 3 we present results for both methods, for the complete data set of 198 introductory statistics students, as well as for a single section of 44 students from this course. These data, collected in 2009, are from students at a Canadian, primarily undergraduate, university.

Table 3: Parameter Estimates using Best Simulation-based Methods for real SATS Data.

	n= 198			**n = 44**		
	Obs	**E$\hat{\rho}$**	**A$\hat{\rho}$**	**Obs**	**E$\hat{\rho}$**	**A$\hat{\rho}$**
$\boldsymbol{\alpha_1}$	0.64	0.82	0.89	0.53	0.69	0.76
$\boldsymbol{\alpha_{gain}}$	-0.36	-0.18	-0.11	-0.47	-0.31	-0.24

From Table 3 we see that all the estimates for the gain slope are negative, suggesting that the negative relationship depicted in Figure 1 is not all due to measurement error. We repeated this approach on several other SATS data sets, looking at five of the SATS components; most of the estimates for the true slope were negative, and the confidence intervals either exclude zero or have an upper bound just above zero. Note that we did not consider the effort component as the data is extremely skewed.

After allowing for regression to the mean in the analysis of SATS data, we estimate the gains are still negatively related to the pre-score. Therefore, we recommend that adjustment be made for measurement error when analyzing of SATS gain scores relative to pre-scores. What we have discovered is that estimates for "true" coefficients using the test-retest reliability coefficients are far superior to those using the correlation between pre- and post-scores. The choice of adjustment method, though, depends on the true slope and the test-retest coefficient. The EIV estimator using the true value of ρ_{xx} is known to be unbiased, as is illustrated in our simulations, and shows only slight bias when using the M-W estimate for ρ_{xx}. Although the adjusted post-score estimator is noticeably biased, it usually has a smaller variance than the EIV estimator. More importantly, both of the two adjustment methods are far superior to using the naïve unadjusted regression results. We are now investigating the impact of regression-to-the-mean in the case where we compare SATS scores for two groups of students to assess an intervention.

Acknowledgements

We would like to thank Candace Schau and Marjorie Bond for founding the CAUSE research cluster in student attitudes to statistics. We appreciate their time and assistance; they are wonderful mentors. We also thank Candace for developing the SATS survey, and for providing data from the SATS database data for estimation of the test-retest reliabilities for the attitude components. We would like to thank Yu ("Maggie") Lin, a graduate student at The University of Western Ontario, for running additional simulations.

The research reported in this paper was supported partially by the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), through NSF Award # 0618790.

References

Fuller, Wayne A. (1987) *Measurement Error Models,* Wiley

Millar, A. M., & Schau, C. (2010). *Assessing students' attitudes: the good, the bad, and the ugly*. Joint Statistical Meetings, Vancouver, http://www.statlit.org/pdf/2010MillarSchauASA.pdf.

Roberts, A. O. H. (1980). "Regression Toward the Mean and the Regression-Effect Bias," in G. Echternacht (ed.), *Measurement Aspects of Title I Evaluations (pp. 59-82), San Francisco: Jossey-Bass.*

Rocconi, L. M., & Ethington, C. A. (2009), "Assessing Longitudinal Change: Adjustment for Regression to the Mean Effects," *Research in Higher Education, 50, 368-376.*

Schau,C. SATS survey :'SATS Scoring' and 'View SATS' www.evaluationandstatistics.com.

Shultz, K. S. and Koshino, H. (1998) "Evidence of reliability and validity for Wise's Attitude Toward Statistics scale" *Psychological Reports*, 82, 27-31.

Tapia, M. and Marsh II, G. E., "An instrument to measure mathematics attitudes" (2004) *Academic Exchange Quarterly*, 8(2), http://www.rapidintellect.com/AEQweb/cho25344l.

Vanhoof, S., Sotos, A. E. C., Onghena, P., Vershaffel, L., Van Dooren, W., Van den Noortgate, W (2006) "Attitudues Toward Statistics and Their Relationship with Short- and Long-Term exam Results" *Journal of Statistics Education* [Online], 14(3) www.amstat.org/publications/jse/v14n3/vanhoof.html.