Nonlinear Modeling with Big Data in SAS® and JMP

Chong Ho Yu, Azusa Pacific University, Azusa, CA

Hyun Seo Lee, Azusa Pacific University, Azusa, CA

Siyan Gan, Pepperdine University, Malibu, CA

Emily Brown, Azusa Pacific University, Azusa, CA

ABSTRACT

Very often conventional statistical procedures, such as Pearson correlation and regression modeling, impose an unrealistic linear structure on the data. However, in the real world many relationships are curvilinear. When the sample size is very large, it is difficult to detect non-linear patterns in the over-plotted data cloud. In this presentation, the data sets of 2015 Programme for International Student Assessment (PISA) and 2016 Programme for International Assessment of Adult Competencies (PIAAC) were utilized to demonstrate how self-efficacy in science and self-report readiness to learn are correlated with test performance in a nonlinear fashion. The SAS® and JMP procedures for detecting these nonlinear patterns include Lambda smoothing spline via interactive data visualization, artificial neural networks, and transformation regression. Merits and weaknesses of each technique are discussed.

INTRODUCTION

George Box (1987) wrote, "All models are wrong but some are useful" (p.424). A linear model might be wrong in the sense that it highly simplifies the phenomenon being studied, yet it could still yield practical applications in some situations (Yu, 2010). No doubt linear modeling is convenient and easier to interpret. For this reason even though some researchers are aware that the underlying relationship between the variables is nonlinear, they transform the data for a linear fit. Whether linear modeling with transformed data or nonlinear modeling is a better analytical strategy is an ongoing debate (Xiao, White, Hooten, & Durham, 2011; Packard, 2014). Indeed, the outcome depends on the criterion chosen, such as the lowest error rate (i.e. fitness) or lowest AIC (i.e. parsimoniousness). However, the criterion of "usefulness" advocated by George Box does not favor one or another. On some occasions, a complicated and well-fitted model is more useful but in other situations a simple model is appropriate.

Based on the advice by George Box the authors argue that in order to produce meaningfully and practical applications in psychology, psychological researchers should refrain from committing themselves to linear modeling (e.g. Pearson's *r*, OLS regression) prematurely. Rather, one has to take social consequences into account while going back and forth between fitness and parsimoniousness. This process is both interactive and iterative, and thus JMP and SAS, which equip users with exploratory tools, are highly suitable to this task. In the following sections, nonlinear modeling will be illustrated with the example of the relationship between individual's self-efficacy and academic performance.

DEBATE IN SELF-EFFICACY

A few decades ago social psychologist Albert Bandura (1990, 1991, 1995, 1997) revolutionized education by stressing the importance of self-efficacy in learning. According to the theory of self-efficacy, one's belief in one's own ability can lead to success in accomplishing a task or fulfilling a goal. Hence, his theory significantly contributed to the development of positive psychology (Bandura, 2008). However, Bandura (2012) also emphasized that self-efficacy is not an omnibus trait that is invariant across all facets of human experiences. Many educators in the U.S. misconstrued the notion of self-efficacy by boosting learners' self-images regardless of their performance. As a result, "everyone gets a trophy" syndrome became a mockery in the US education. Specifically, U.S. schools are flooded with grade inflation, as evidenced by the fact that today 43 percent of all U.S. college students receive "A" grades, whereas 86% of private school students receive "B" grades or higher (Sigman, 2012). Twenge (2010) argued that the 'everybody gets a trophy' mentality fails to build true self-esteem; instead, it leads to ego inflation and a sense of entitlement. Not surprisingly, some researchers found that contrary to the promised results, boosting self-efficacy is debilitating to actual performance (Vancouver & Kendall, 2006; Vancouver, More, & Yoder, 2008; Vancouver, Thompson, Tischner, & Putka, 2002; Vancouver, Thompson, & Williams, 2001). Nevertheless, one should not blame Bandura for this unintended consequence. Bandura (2012) himself had explicitly warned people of the danger of

creating a false sense of self-efficacy "If people experience only easy successes they come to expect quick results and are easily discouraged by setbacks and failures. Resilient self-efficacy requires experience in overcoming obstacles through perseverant effort. Resilience is also built by learning how to manage failure so that it is informative rather than demoralizing" (p.13).

As mentioned before, Vancouver and his colleagues found an inversed relationship between one's self-efficacy and performance. Nonetheless, some researchers reached the opposite conclusion by utilizing linear regression or/and structural equating modeling (Chevaliera, Gibbons, Thorpee, Snelle, & Hoskinsf, 2009; March & Martin, 2010; Morony, Kleitman, Lee, & Stankov, 2013; Williams & Williams, 2010). Both camps might be right. In one part of a curvilinear relationship, the association between self-efficacy and performance is positive. However, after the inflection point, the relationship is reversed. If educators focus on one part of the curve, the model might fail to provide useful guidelines for education reform. In the following section the data sets of 2015 Programme for International Student Assessment (PISA) and 2016 Programme for International Assessment of Adult Competency (PIAAC) were utilized to demonstrate how one's self-efficacy in science and self-reported readiness to learn are correlated with test performance in a nonlinear fashion. Uncertainty is an inherent property of any estimation and thus for every participant OECD reported ten plausible test scores, known as plausible values (PV) (OECD, 2013). Wu (2004) suggested randomly selecting PVs for analysis because each PV is equally good for estimating the population characteristics, as each set forms a sampling distribution of the estimated student ability. Following the advice of Wu, this study used a randomly selected PV for both PISA and PIAAC.

EXAMPLE 1: PISA 2015

Programme for International Student Assessment (PISA), as the name implies, is a large-scale cross-cultural assessment of student performance administered by Organization for Economic Cooperation and Development (OECD) every four years. The target population of this assessment consists of 15-year old school children in 65 member nations of OECD. The assessment covers literacy, math, and science, with a different focus in each round. In 2015 the focus was on science learning and therefore data of science self-efficacy (SSE), science self-belief (SSB), and ambition in general were collected. Science self-efficacy refers to a student's perception of his or her ability to successfully complete a science-related academic task or reach an academic goal in science. Self-belief, also known as self-concept, refers to the self-evaluation of one's general ability in a domain (Marsh & Martin, 2011). Ambition is measured by the degree to the statement "I see myself as an ambitious person." Table 1 is a summary of average math and science test scores, science self-efficacy scores, self-belief scores, and ambition scores of USA and seven Asian countries/regions.

Table 1. Average 2015 PISA scores of USA and Asian countries/regions.

Country/Region	Math scores		Science scores		Self-efficacy		Self-belief		Ambition	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
China	541.74	100.84	528.34	98.56	0.06	1.17	0.19	0.87	2.98	0.73
Hong Kong	550.55	88.48	525.60	79.58	-0.07	1.22	0.21	0.95	2.80	0.80
Japan	533.30	88.18	539.03	93.28	-0.46	1.22	-0.51	1.02	2.64	0.82
Macau	543.98	79.03	528.59	81.84	-0.03	1.12	-0.50	0.81	2.63	0.81
S Korea	523.91	99.97	514.75	95.00	-0.02	1.23	0.34	0.98	2.84	0.75
Singapore	557.08	95.75	545.95	104.60	0.07	1.14	0.42	0.94	3.00	0.79
Taiwan	539.20	103.79	530.85	99.85	0.19	1.19	-0.01	0.89	2.92	0.76
USA	474.35	87.92	502.60	98.04	0.29	1.29	0.65	0.95	3.25	0.72

The preceding table indicates that students in the U.S. have the lowest math and science scores, in comparison to their peers in Asia. However, their self-efficacy in science, self-belief, and ambition scores are higher than students in Asian countries. At first glance, one might draw the conclusion that boosting one's self-image is detrimental to one's

academic performance because the majority of societies in Asia that accept modesty as the norm outperform the US in PISA. However, a careful researcher should refrain from jumping into a hasty conclusion before thoroughly investigating the data. Due to space constraints, in the following only the relationship between science test scores and science self-efficacy is used for illustration.

Detecting patterns with a big data set is challenging. In this data set there are 54,978 observations, and as a result the scatterplot of science scores and science efficacy is jammed with data points (see Figure 1). When an OLS regression is run, the regression line is perfectly straight, of course, yet, misleading.

Figure 1. OLS regression model of science scores and science self-efficacy.

LAMBDA SMOOTHING

In JMP, there are many different ways to detect the true relationships between variables. The authors advise against using static approaches. For example, in **Fit Curve** under **Specialized modeling,** one can choose from a plethora of nonlinear fitting techniques. However, counting on numeric criteria (e.g. AICc, BIC, SSE, R-square…etc.) could result in an erroneous conclusion. To remediate the situation, one can utilize the **Lambda smoother** in **Graphic Builder** to acquire a holistic view of the data visually and interactively. By default, the Lambda smoothing parameter is set to the middle, as shown in Figure 2. Under this configuration the model is too complicated to be useful because the regression curves have too many turns (ups and downs).

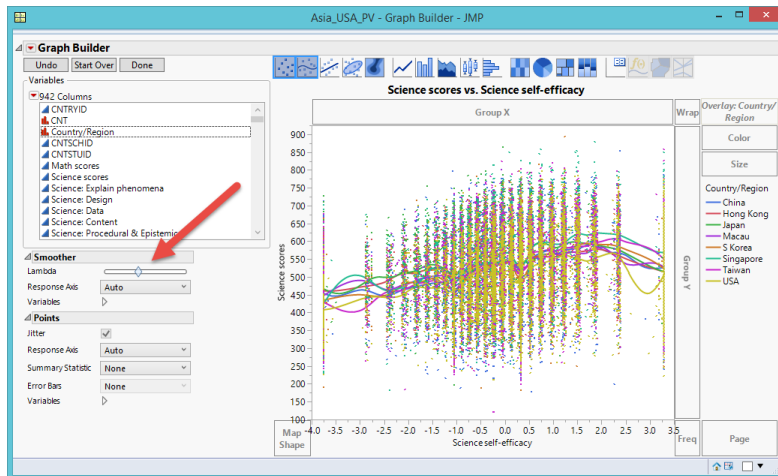**Figure 2. Over-complicated nonlinear model.**

The analyst could go to another extreme by pushing the Lambda parameter to the rightmost polar, as displayed in Figure 3. In this case the regression lines of all countries are perfectly straight, but this model is too simplistic to be useful.
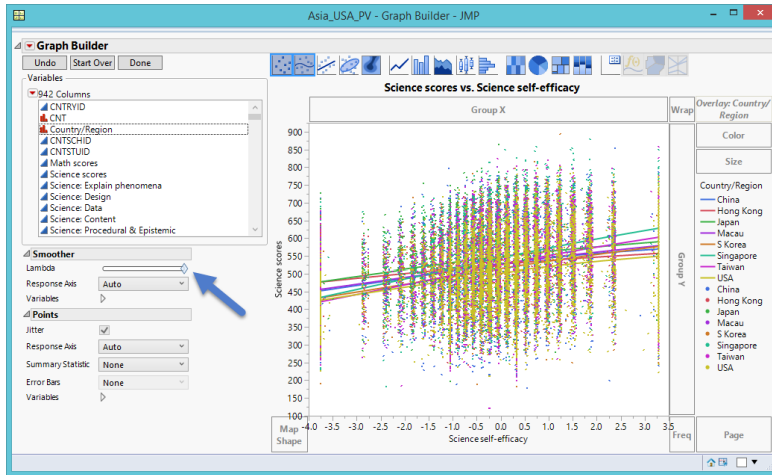

Figure 3. Over-simplified linear model.

By a simple adjustment in the interactive smoother, the analyst could detect a curvilinear relationship between science test performance and science self-efficacy (see Figure 4). It seems that Bandura was right. As self-efficacy improves, science test scores increase. However, Twenge and Vancouver were right as well. After science self-efficacy passes a certain threshold, science test performance is reduced. It is important to point out that this nonlinear pattern is not only a phenomenon confined to students in America. This also occurs in Asian countries/regions. This model is considered more useful in two ways: it explains why educators implementing Bandura's theory did not produce expected results and also necessitates a reform in our pedagogical approaches.
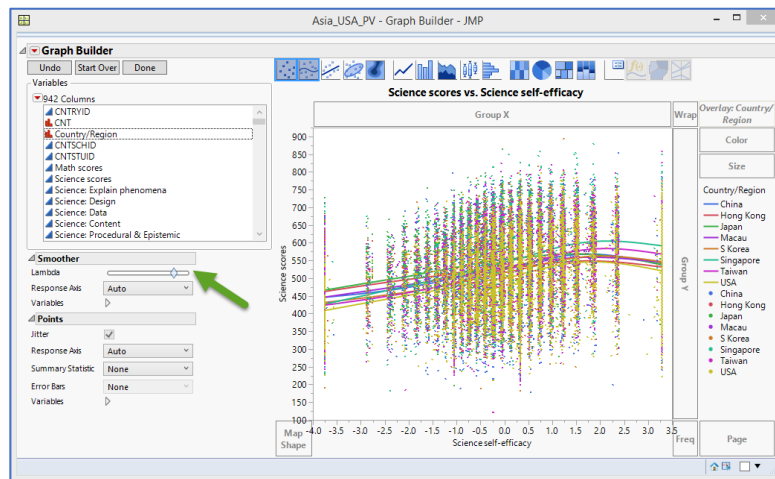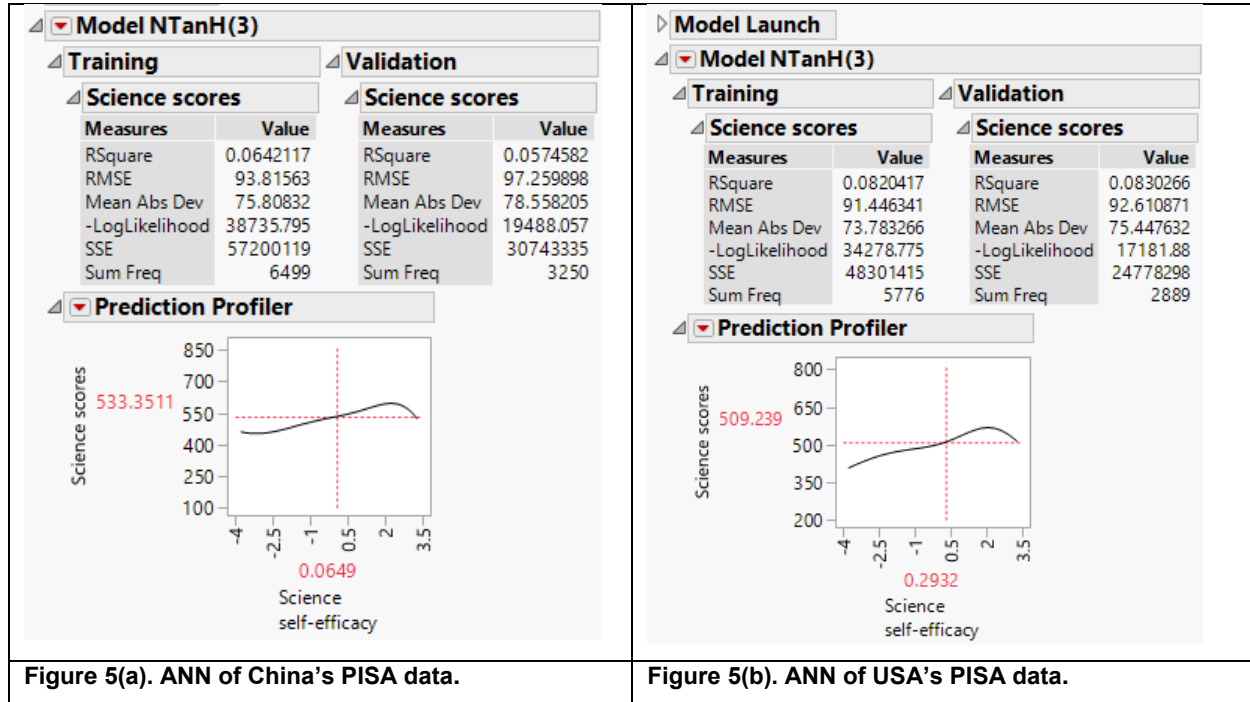

Figure 4. Useful nonlinear model.

ARTIFICAL NEURAL NETWORK

Traditional journals might be skeptical of visual reports and exploratory data analysis. To rectify this situation, the analyst could employ artificial neural networks (ANN) for nonlinear modeling, which report both visual displays and statistical figures. Neural networks, as the name implies, try to mimic interconnected neurons in animal brains in order to make the algorithm capable of complex learning for extracting patterns and detecting trends (Kuan, 1994;

McMenamin, 1997). As this approach artificially mimics human neurons in computers, it is also named artificial neural networks. ANN is a form of machine learning, meaning that to a large degree the artificial intelligence algorithm, not the analyst, makes the decision. It is built upon the premise that real world data structures are complex and nonlinear, and thus it necessitates complex learning systems. Unlike regression modeling that assumes linearity, neural networks could model linearity and thus they typically outperformed regression (Somers & Casal, 2009). A trained neural network can be viewed as an "expert" in the category of information it has been given to analyze. This expert system can provide projections given new solutions to a problem and answer "what if" questions. A typical neural network is composed of three types of layers, namely, the input layer, hidden layer, and output layer. To save space only the ANN results of China and USA are shown below. One may argue that manual smoothing of nonlinear curves is subjective, but ANN, which is capable of learning from the data, also suggests that while enhancing self-efficacy can improve performance initially, inflated one's self-efficacy could drag down one's actual performance. This pattern is common across cultures.

Figure 5(a). ANN of China's PISA data.	Figure 5(b). ANN of USA's PISA data.

One major criticism against ANN is that it acts like a black box, forcing humans to hand over judgment to the machine while we don't fully understand what is happening inside the process.

TRANSFORMATION REGRESSION

To alleviate this problem, the analyst can program nonlinear models in SAS and make most of the decisions. Two main nonlinear modeling methods exist in SAS, namely, NLIN (nonlinear modeling) and TRANSREG (transformation regression). NLIN is the extreme opposite end to ANN: unlike ANNA that is highly automated, NLIN requires the analyst to specify the initial values of parameters and model equations, and thus making programming intimating. This article focuses on TRANSREG, which is more user-friendly. TRANSREG fits models by providing users with many transformation options as follows:

- Non-optimal transformations: e.g. ARS (inverse trigonometric sine transformation), EXP (exponential)

- Nonlinear transformations: e.g. PBS (penalized B-spline transformation, Eilers & Marx, 1996), SMO (smoothing spline transformation, Reinsch, 1967).

- Optimal transformations: e.g. MSP (B-spline transformation with monotonic coefficients, De Boor, 1978; De Leeuw, 1986).

Although there are over 20 transformation options, this illustration focuses on three only: Smoothing Spine, Penalized B-spine, and Monotonic Spine. To save space, only the Japan sample is discussed here. The SAS code for generating various transformation regression models is shown next:

```
ods graphics on;
proc transreg; model identity(science_scores) = SMO(Science_self_efficacy);
proc transreg; model identity(science_scores) = PBS(Science_self_efficacy);
proc transreg; model identity(science_scores) = MSPLINE(Science_self_efficacy);
run;
```

The dependent variable "science scores" is specified with the syntax IDENTITY, meaning that "science scores" is analyzed as is, without transformations. Figure 6 shows the result of the Smoothing Spine fit. Smoothing is intended to balance the goodness of fit and smoothness with a noisy data set. However, it appears that there are exceeding numbers of knots (turn points) in this model; the line goes up and down drastically. Hence, this overly complicated and noisy model does not seem to be useful.
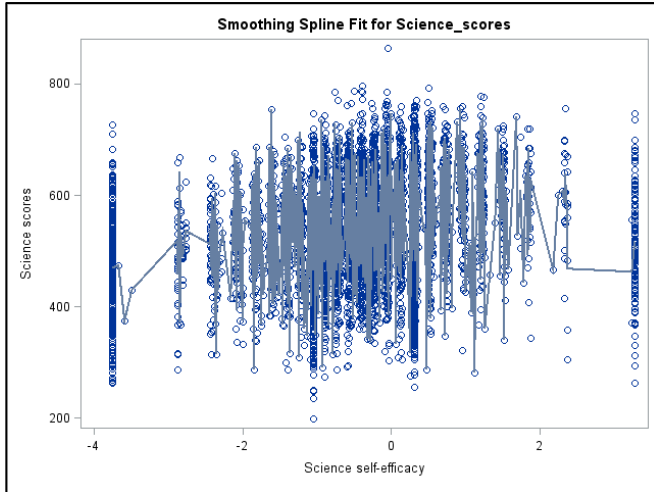

Figure 6. Smoothing Spline model.

Figure 7 shows the result of the Penalized B-Spine fit. The main idea of PBS is to penalize complexity while controlling the smoothness of fit. It yields a smoother curve than Smoothing Spline because it pre-determines the location of knots evenly works for most observations. By default, PBS is fit with 100 evenly spaced knots, three evenly spaced exterior knots, and a difference matrix of order three.
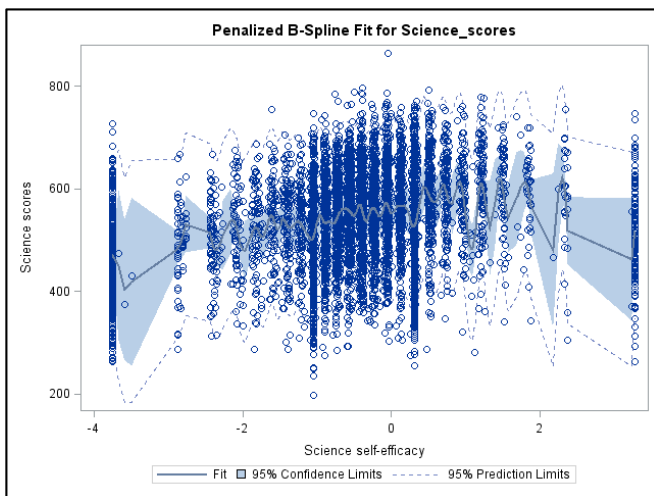

Figure 7. Penalized B-spline model.

6

Figure 8 displays the result of Monotonic Spline fit, which is the smoothest of all three TRANSREG models. At first glance, it is similar to the output produced in JMP. However, not only the curve is over-smoothed but the low degree of steepness of the curve suggests a minimal association between science test performance and SSE. Hence, this model might be too simplistic to be useful. On the other hand, the PBS model is comparable to the JMP results. In addition, it provides the analyst with more details, such as local fluctuations of the curve and confidence intervals. By looking at Figure 7, one can reach the same conclusion: positive self-efficacy could boost test performance to some extent, but inflated ego could lead to the opposite outcome.
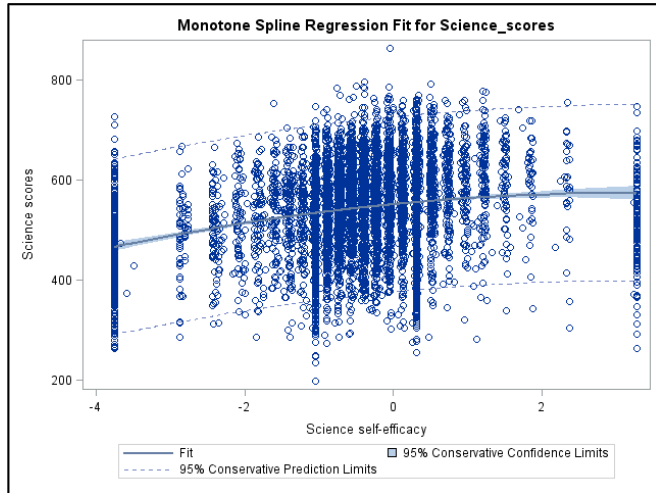

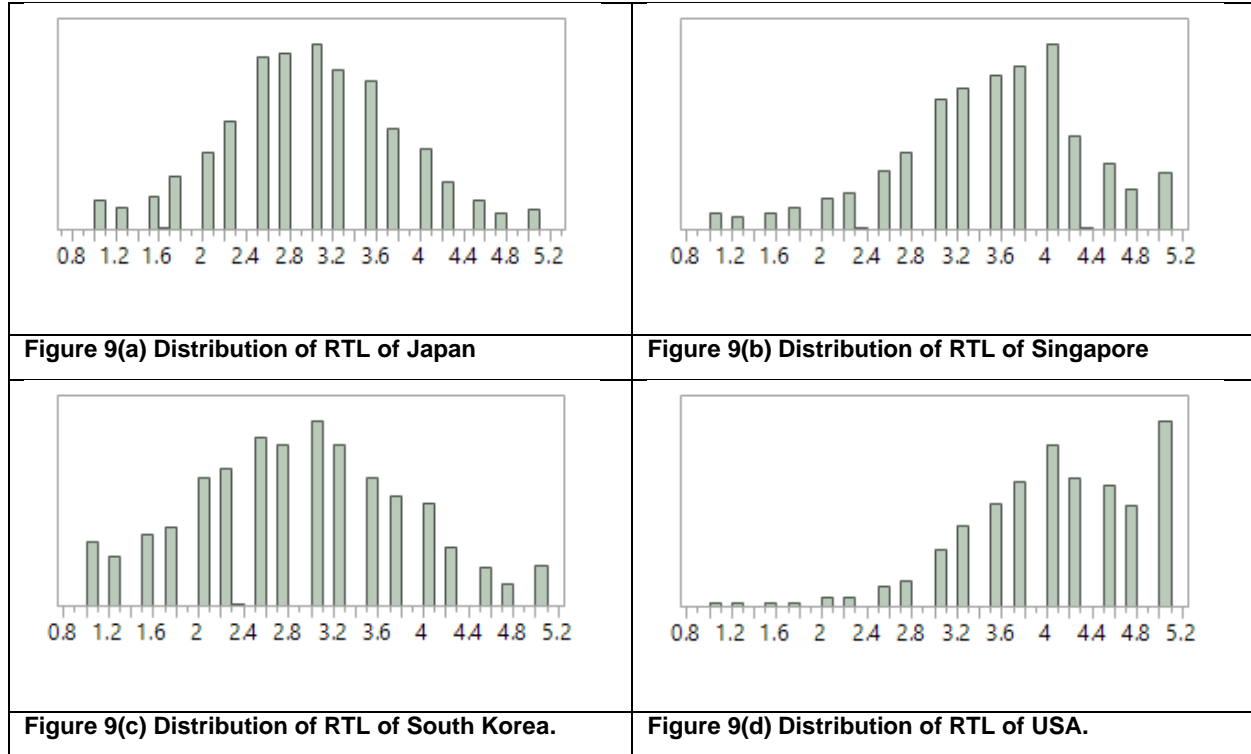**Figure 8. Monotone Spline model.**

EXAMPLE 2: PIAAC 2016

Programme for International Assessment of Adult Competency (PIAAC) is an adult version of PISA. The age range of PIAAC participants is between 16 and 64. In PIAAC three skill domains, including literacy, numeracy, and technology-based problem solving, are assessed. Numeracy and digital problem-solving have become the focal point of US educators because like the PISA results, the US participants also fall behind their international peers in math- and science-related skills (OECD, 2012). This may imply that the US workforce might be less competitive in today's technology-driven economy. In addition to test scores, PIAAC also collected data regarding self-reported readiness to learn. Readiness to Learn (RTL) is defined as the tendency to learn new things, relate information to prior knowledge and life situations, as well as engage in problem-solving and information-seeking behaviors. Table 2 is a summary of the average numeracy scores, digital problem-solving scores, and readiness to learn scores of the US and three Asian countries.

Table 2. Average 2015 PIAAC scores of USA and Asian countries (n = 22,423).

	Numeracy		Problem-solving		Readiness to learn	
Country	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Japan	289.47	43.82	295.31	43.77	2.96	0.81
Singapore	255.81	69.23	287.66	45.01	3.48	0.83
South Korea	262.28	46.14	283.19	37.24	2.86	0.94
United Sates	254.51	55.56	278.01	43.28	4.05	0.72

As shown in Table 2, American participants report the highest level of readiness to learn, but test performance in numeracy and digital problem-solving is lower than their counterparts in Asian countries. Figure 5 is even more revealing. The RTL distributions of Japan and South Korea are fairly normal while that of Singapore is slightly skewed. However, not only the US distribution is skewed towards high RTL values, but also the mode is 5 (the highest possible value), meaning that the vast majority of the U.S participants view themselves as being extremely ready to learn.

Figure 9(a) Distribution of RTL of Japan

Figure 9(b) Distribution of RTL of Singapore

Figure 9(c) Distribution of RTL of South Korea.

Figure 9(d) Distribution of RTL of USA.

LAMBDA SMOOTHING

As aforementioned, when the relationship between numeracy and RTL is investigated through Lambda smoother, one can see a curvilinear association: Nonlinear patterns are detected across cultures. Numeracy scores improve as RTL goes up, but after passing the inflection point, the line turns flat or goes down slightly. However, unlike inflated self-efficacy that could adversely affect performance, increasing perceived RTL has less adverse effect on individual's performance.
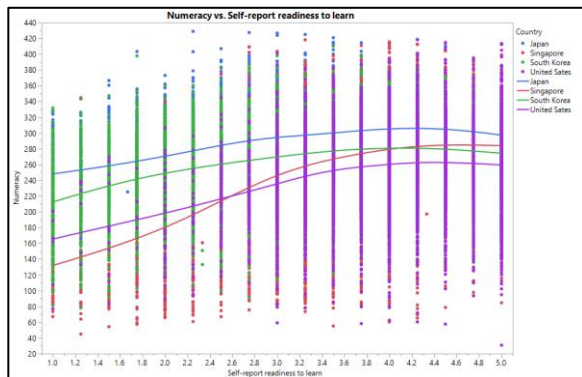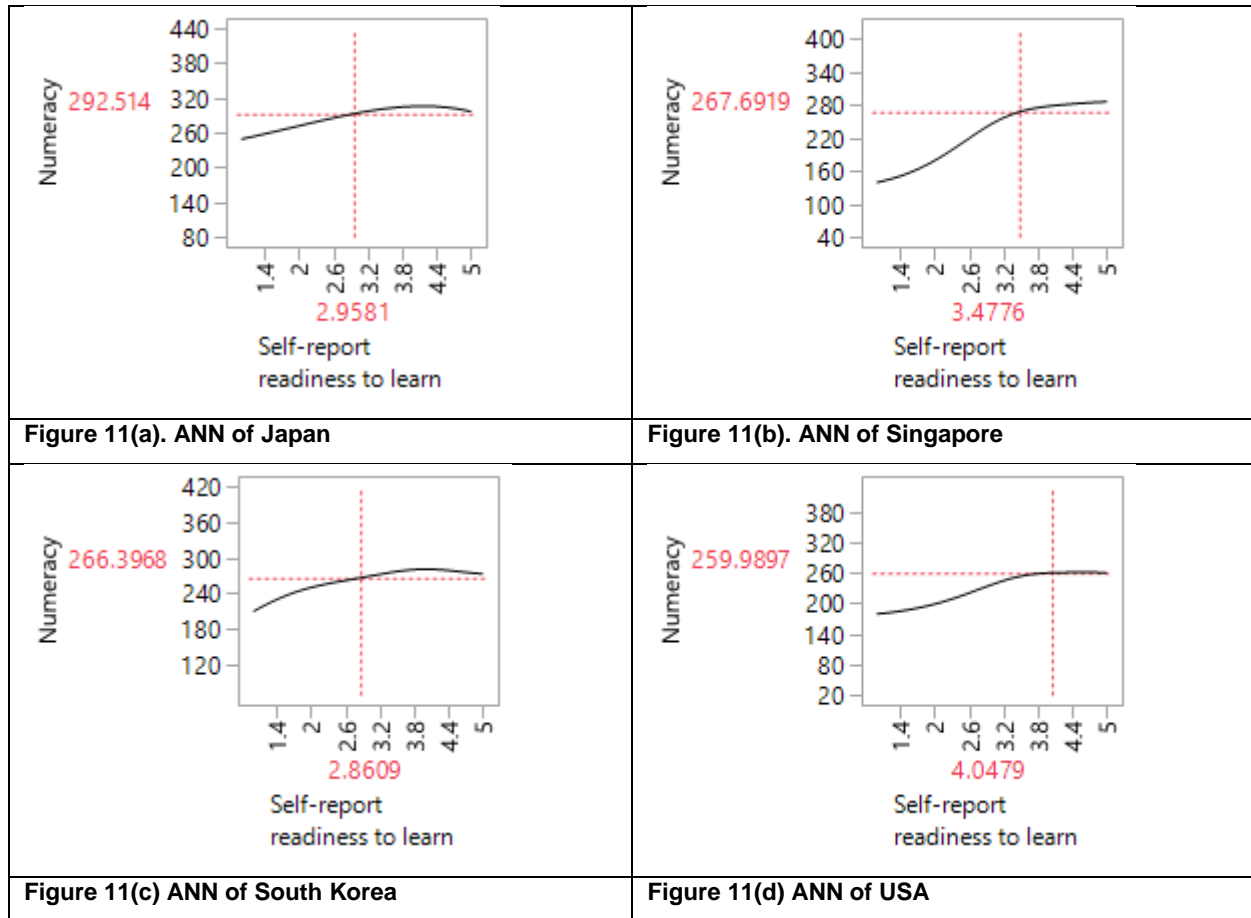

Figure 10. Useful nonlinear model by Lambda smoothing

ARTIFICAL NEURAL NETWORK

ANN was employed to verify the results. The findings of Lambda smoothing and ANN were congruent to each other. Once again, self-perception of readiness to learn and math skills go up hand in hand to a certain point. However, once it passes the threshold, one's inflated ego could no longer be helpful in learning or even drag down performance.

Figure 11(a). ANN of Japan
Figure 11(b). ANN of Singapore
Figure 11(c) ANN of South Korea |
Figure 11(d) ANN of USA |

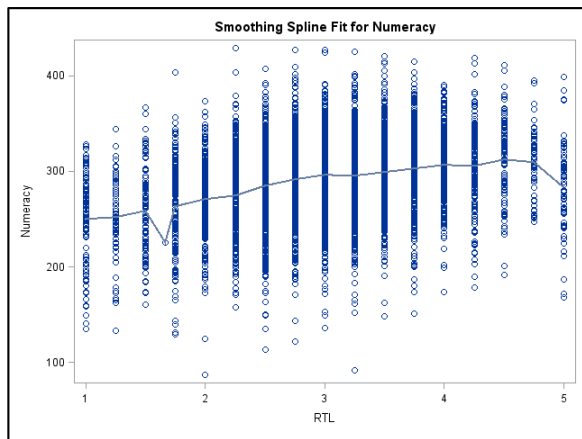## TRANSFORMATION REGRESSION
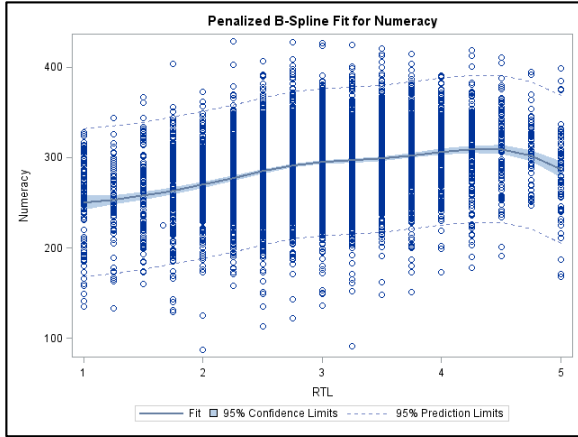

Figure 12. Smoothing Spline model.

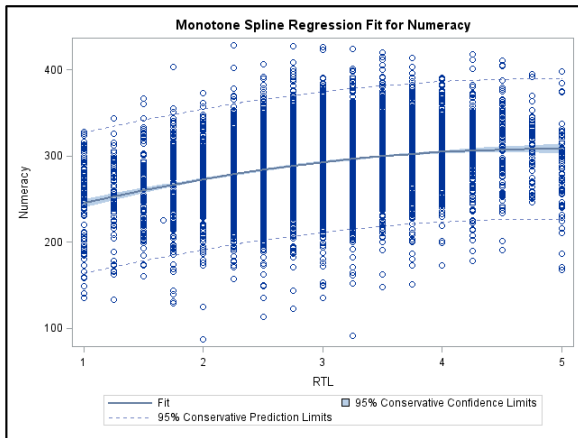**Figure 13. Penalized B-Spline model.**

Figure 14. Monotonic Smoothing model.

Due to space constraints, only the Japan sample was used for computing transformation regression models. Interestingly, this time Smoothing Spline, PBS, and Monotonic Spine yielded similar results (Figure 12-14). However, Smoothing Spline has a strange depression at the lower left of the graph while Monotonic Spine does not show any down-turn at all. Hence, the PBS model is considered the most informative and useful model.

CONCLUSION

Linear modeling is convenient, but it often fails to capture the essence of reality, resulting in misguided decisions. Whether boosting self-efficacy is the key to one's better learning has become a pressing issue in the field of education. However, the answer should not be dichotomous (Yes, it can/No, it cannot). It seems that previous research overlooked the nonlinear character of this association, fueling unnecessary debates. This article demonstrates how different tool sets in JMP and SAS can be utilized to identify a useful model by detecting nonlinear relationships in data sets. When the sample size is extremely large, it is very difficult to optically examine the data structure due to over-plotting. As a remedy, interactive and exploratory Lambda smoothing can be helpful to unmask nonlinear patterns. However, this method heavily relies on subjective judgment. To compensate for the possible risk of misjudgment, artificial neural networks can be employed to suggest a nonlinear model by providing an advantage of fully automated machine learning approach. When the analyst would like to seek for the middle ground between the two extremes (highly subjective human decision vs. automated machine learning), he or she can experiment with different smoothing options in TRANSREG.

REFERENCES

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1990). Reflections on notability determinants of competence. In R. J. Sternberg & J. Kolligian Jr. (Eds.), *Competence considered* (pp. 315-362). New Haven, CT: Yale University Press.

Bandura, A. (1991). Self-regulation of motivation through anticipatory and self-regulatory mechanisms. In R. A. Dienstbier (Ed.), *Perspectives on motivation: Nebraska symposium on motivation*, Vol. 38 (pp. 69-164). Lincoln, NE: University of Nebraska Press.

Bandura, A. (Ed.). (1995). *Self-efficacy in changing societies*. Cambridge, UK: Cambridge University Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Bandura, A. (2008). An agentic perspective on positive psychology. In Shane J Lopez (Ed.). *Positive psychology* (pp. 167-196). Westport, CT: Greenwood Publishing.

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management, 38*, 9-44. DOI: 10.1177/0149206311410606

Box, G. E. P., and Draper, N. R., (1987). *Empirical model building and response surfaces*. New York, NY: John Wiley & Sons.

Chevaliera, A., Gibbons, S., Thorpe, A., Snell, M., & Hoskins, S. (2009). Students' academic self-perception. *Economics of Education Review, 28*, 716-727.

De Boor, C. (1978). *A practical guide to splines*. New York, NY: Springer-Verlag.

De Leeuw, J. (1986). *Regression with optimal scaling of the dependent variable*. Leiden, Netherlands: Department of Data Theory, University of Leiden.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89–121.

Kuan, C., & White, H. (1994). Artificial neural networks: An econometric perspective. *Econometric reviews, 13*, 1-91.

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*, 59–77. doi:10.1348/000709910X503501.

McMenamin, J. S. (1997). A primer on neural networks for forecasting. *Journal of Business Forecasting, 16*, 17–22.

Morony, S., Kleitman, S., Lee, Y. M., & Stankov, L. (2013). Predicting achievement: Confidence vs self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research, 58,* 79-96.

Organization for Economic Co-operation and Development [OECD]. (2012). *Literacy, numeracy and problem solving in technology-rich environments framework for the OECD survey of Adult Skill*. DOI: 10.1787/9789264128859-en. Retrieved from http://www.oecd.org/skills/piaac/PIAAC%20Framework%202012--%20Revised%2028oct2013_ebook.pdf

Organization for Economic Co-operation and Development [OECD]. (2013). *Technical report of the survey of adult skills (PIAAC).* Retrieved from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

Packard, G. (2014). Review article: On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Biological Journal of the Linnean Society, 113*, 1167–1178.

Reinsch, C. H. (1967). Smoothing by Spline Functions. *Numerische Mathematik, 10*,177–183.

Sigman, M. (2012, June 18). When everyone gets a trophy, no one wins. *Huffpost*. Retrieved from http://www.huffingtonpost.com/michael-sigman/when-everyone-gets-a-trop_b_1431319.html

Twenge, J. M. (2010). *The Narcissism epidemic: Living in the age of entitlement*. New York, NY: Atria Books.

Vancouver, J. B., & Kendall, L. N. 2006. When self-efficacy negatively relates to motivation and performance in a learning context. *Journal of Applied Psychology, 91,* 1146-1153.

Vancouver, J. B., More, K. M., & Yoder, R. J. 2008. Self-efficacy and resource allocation: Support for a nonmontomic discontinuous model. *Journal of Applied Psychology, 93*, 35-47.

Vancouver, J. B., Thompson, C. M., Tischner, E. C., & Putka, D. J. 2002. Two studies examining the negative effect of self-efficacy on performance. *Journal of Applied Psychology, 87*, 506-516.

Vancouver, J. B., Thompson, C. M., & Williams, A. A. 2001. The changing signs in the relationships between self-efficacy, personal goals, and performance. *Journal of Applied Psychology, 86*, 605-620.

Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology, 102*, 453-466. DOI: 10.1037/a0017271

Xiao, X, White, E. P., Hooten, M. B., & Durham, S. L. (2011). On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Ecology, 92*, 1887-1894.

Wu, M. (2004). Plausible values. *Rasch Measurement Transactions,18*, 976-978.

Yu, C. H. (2010). A model must be wrong to be useful: The role of linear modeling and false assumptions in theoretical explanation. *Open Statistics and Probability Journal, 2*, 1-8. Retrieved from http://benthamopen.com/ABSTRACT/TOSPJ-2-1

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chong Ho Yu
Azusa Pacific University
901 E. Alosta Ave.
Azusa, CA 91702
cyu@apu.edu
http://www.creative-wisdom.com