Optimal Data Analysis
Vol. 8 (January 15, 2019), 15-20

2155-0182/10/$3.00

When to Evaluate a Nonlinear Model

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Two bivariate data sets and four regression models provide the solution.

This paper is organized in consecutive sections in which—in reality, either a nonlinear or linear *relationship* underlies two variables, X and Y.

Both of the sections present consecutive findings of nonlinear and linear *models* of the relationship between X and Y.

Table 1 presents the artificial data used in this simulation study.

Table 1: Simulated Data for Exposition

Nonlinear Relationship		*Linear* Relationship	
X	Y	X	Y
10	50	10	50
12	32	11	48
20	20	18	46
25	12	26	38
30	11	34	32
35	16	38	30
40	21	44	26
50	34	48	22

Underlying *Nonlinear* Relationship

Nonlinear Model

As seen in Figure 1, data indicate one inflection point: the highest recorded value of Y (vertical axis) occurs at the lowest recorded value of X (horizontal axis); Y decreases monotonically until reaching X=30 and then reverses, increasing monotonically until X reaches its highest recorded level. In geometric terms, Y follows a parabolic path over X.

The number of inflections in Y which occur over the domain of X indicates the terms to include in the model: no inflections indicate a *linear* model with an intercept and a coefficient for X; one inflection point indicates a *nonlinear* model with an additional coefficient for X^2; two inflections indicate a *nonlinear* model with an additional coefficient for X^3; etcetera.

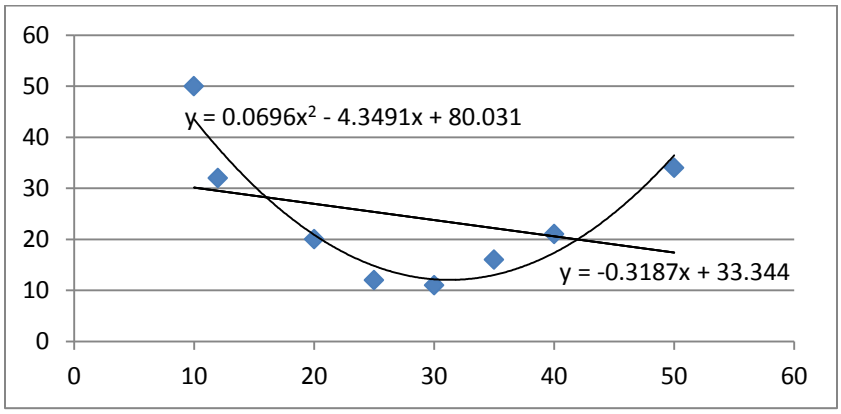
Visually the nonlinear model seems to be an excellent representation of the actual data. For example, to predict Y for X=30 by the equation in Figure 1, Y-Hat=0.0696*30^2–4.3491*30+80.031=12.2, a relatively close estimate of the true value 11 (Table 1). The omnibus model is statistically significant [F(2,5)=24.53, $p<0.0026$] and accurate (R^2=0.908), and both the linear (t=-6.93, $p<0.001$) and nonlinear (t=6.57, $p<0.0012$) coefficients were significant.

Linear Model

Visually the linear regression model appears as a poor representation of the data—only two data points lie in close proximity to the model. The model is not statistically significant [F(1,6)= 0.74, $p<0.43$] and very weak (R^2=0.109)—only capable of accurately predicting the mean.[1]

2155-0182/10/$3.00

Figure 1: Nonlinear Data, Linear *vs*. Nonlinear Models



Underlying *Linear* Relationship

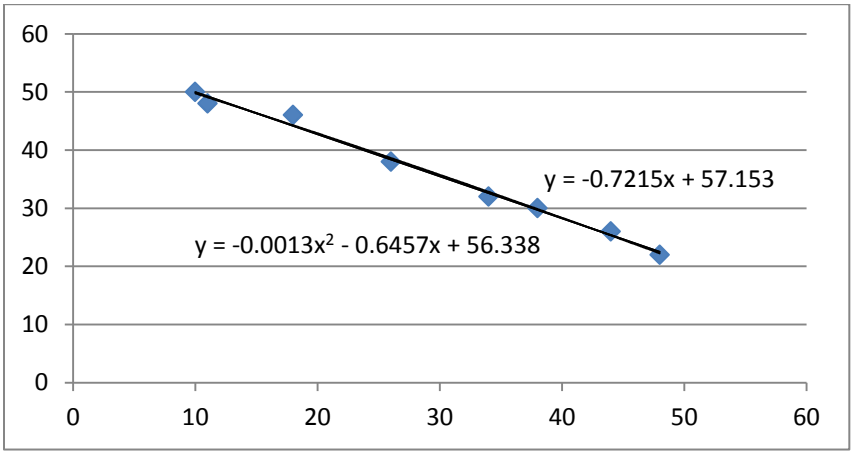
Nonlinear Model

As seen in Figure 2, data indicate what visually appears to be a nearly perfect linear effect. To predict Y for X=34 by the equation given in Figure 2, Y-Hat=-0.0013*34^2–0.6457*34+ 56.338=32.9, a relatively accurate estimate of the true value of 32 (Table 1). The omnibus model is significant [F(2,5)=337.4, p<0.0001] and accurate (R^2=0.993): although the linear coefficient was significant (t=-4.06, p<0.0098), the nonlinear coefficient was not statistically significant (p<0.65).

Linear Model

The model is statistically significant [F(1,6)= 773.33, p<0.0001] and accurate (R^2=0.992). Predicting Y for X=34 by the equation given in Figure 2, Y-Hat=-0.7215*34+57.153=32.6, a relatively accurate estimate of the true value of 32 (Table 1).

Figure 2: Linear Data, Linear *vs*. Nonlinear Models



2155-0182/10/$3.00

Comments

Results showed that if a nonlinear effect exists then a *nonlinear* model *will* find the nonlinear effect, but a *linear* model *won't* find the nonlinear effect. And, if a linear effect exists then both *linear* and *nonlinear* models *will* find the linear effect (the nonlinear model may or may not also find a nonlinear effect). This suggests it is never a bad idea to examine possible nonlinearity in one's data. Other than finding a perfect linear model, it is difficult to imagine a good reason to not check for nonlinearity in data.

Simulated data used here reflected close to perfect effects, not uncommonly obtained in engineering-related applications, but less often realized in most empirical science. Systematic manipulation of small synthetic data sets such as those used here will familiarize researchers new to nonlinear modeling with the effect of changes in data configurations upon changes in *p*-values, model coefficients, R^2, etcetera, for such general linear models. Conceptually-related research is needed to study the nonlinear performance of all linear statistical methodologies.[2-5] The finding of statistically significant linear *and* nonlinear terms for the nonlinear model when it is applied to the nonlinear data suggests a possible synergy with Markov analysis research investigating the number of processes underlying a transition table, as well as heterogeneous populations.[4-12]

It is important to recall that predictive accuracy achieved by suboptimal linear models is explicitly *optimized* using ODA to maximize either ESS or PAC (i.e., Percentage Accurate Classification).[13-27] Arguably more important, recall that explicitly *optimal* linear MultiODA models surpass the accuracy achieved by using optimized suboptimal models, and can identify models in applications for which suboptimal linear methods can find nothing.[28-33]

Perhaps it is most important to consider how many of the phenomena studied in science or events experienced in life have simple linear underpinnings? Most currently, development of a family of nonlinear optimal classification tree analysis algorithms, including novometric theory, explicitly identify the most accurate and parsimonious (non)linear models possible for a sample of data.[34-57]

References

[1]Yarnold PR, Bryant FB, Soltysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression *away from* the mean. *Optimal Data Analysis*, *2*, 19-25.

[2]Grimm LG, Yarnold PR (Eds.). *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books, 1995.

[3]Grimm LG, Yarnold PR (Eds.). *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books, 1995.

[4]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC, APA Books.

[5]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[6]Yarnold PR (2016). GenODA structural decomposition *vs*. log-linear model of one-step Markov transition data: Stability and change in male geographic mobility in 1944-1951 and 1951-1953. *Optimal Data Analysis*, *5*, 213-215.

[7]Yarnold PR (2017). Novometric analysis of transition matrices to ascertain Markovian order. *Optimal Data Analysis*, *6*, 5-8.

[8]Yarnold PR (2017). Novometric comparison of Markov transition matrices for heterogeneous populations. *Optimal Data Analysis*, *6*, 9-12.

[9]Yarnold PR (2018). Using ODA to confirm a first order Markov steady state process. *Optimal Data Analysis*, *7*, 72-73.

2155-0182/10/$3.00

[10]Yarnold PR (2018). Using ODA to determine if a Markov transition process is second order. *Optimal Data Analysis, 7*, 74-75.

[11]Yarnold PR (2018). Using ODA to ascertain if stratification yielded different transition matrices. *Optimal Data Analysis*, *7*, 76-77.

[12]Yarnold PR (2019). Maximum-Precision Markov Transition Table: Successive Daily Change in Closing Price of a Utility Stock. *Optimal Data Analysis*, *8*, 3-10.

[13]Linden A, Yarnold PR (2018). Identifying maximum-accuracy cut-points for diagnostic indexes via ODA. *Optimal Data Analysis*, *7*, 59-65.

[14]Linden A, Yarnold PR (2018). Comparative accuracy of a diagnostic index modeled using (optimized) regression *vs*. novometrics. *Optimal Data Analysis, 7*, 66-71.

[15]Yarnold PR, Soltysik RC (1991). Refining two-group multivariable classification models using univariate optimal discriminant analysis. *Decision Sciences*, *22*, 1158-1164.

[16]Yarnold PR, Hart LA, Soltysik RC (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analyses. *Educational and Psychological Measurement*, *54*, 73-85.

[17]Yarnold PR, Bryant FB (1994). A measurement model of the Type A Self-Rating Inventory. *Journal of Personality Assessment, 62*, 102-115.

[18]Weinfurt KP, Bush PJ (1995). Peer assessment of early adolescents solicited to participate in drug trafficking: A longitudinal model. *Journal of Applied Social Psychology*, *25*, 2141-2157.

[19]Yarnold PR, Stille FC, Martin GJ (1995). Cross-sectional psychometric assessment of the Functional Status Questionnaire: Use with geriatric versus nongeriatric ambulatory medical patients. *International Journal of Psychiatry in Medicine*, *25*, 305-317.

[20]Yarnold PR, Bryant FB, Nightingale SD, Martin GJ (1996). Assessing physician empathy using the Interpersonal Reactivity Index: A measurement model and cross-sectional analysis. *Psychology, Health, and Medicine*, *1*, 207-221.

[21]Yarnold BM, Yarnold PR (2010). Maximizing the accuracy of Probit models via UniODA. *Optimal Data Analysis*, *1*, 41-42.

[22]Yarnold PR (2013). Maximum-accuracy multiple regression analysis: Influence of registration on overall satisfaction ratings of emergency room patients. *Optimal Data Analysis*, *2*, 72-75.

[23]Yarnold PR (2013). Assessing technician, nurse, and doctor ratings as predictors of overall satisfaction ratings of Emergency Room patients: A maximum-accuracy multiple regression analysis. *Optimal Data Analysis*, *2*, 76-85.

[24]Yarnold PR (2013). Creating a data set with SAS™ and maximizing ESS of a multiple regression analysis model for a Likert-type dependent variable using UniODA and MegaODA software. *Optimal Data Analysis, 2*, 191-193.

[25]Yarnold PR (2015). Maximizing ESS of regression models in applications with dependent measures with domains exceeding ten values. *Optimal Data Analysis*, *4*, 12-13.

[26]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[27]Yarnold PR (2018). Alternative prediction-interval scaling strategies for regression models. *Optimal Data Analysis, 7*, 44-45.

[28]Yarnold PR, Soltysik RC, Martin GJ (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of

2155-0182/10/$3.00

multivariable optimal discriminant analysis. *Statistics in Medicine, 13*, 1015-1021.

[29]Soltysik RC, Yarnold PR (1994). The Warmack-Gonzalez algorithm for linear two-category multivariable optimal discriminant analysis. *Computers and Operations Research, 21*, 735-745.

[30]Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine, 10*, 601-606.

[31]Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine, 17*, 2405-2414.

[32]Bennett CL, Kim B, Zakarija A, Bandarenko N, Pandey DK, Buffie CG, McKoy JM, Tevar AD, Cursio JF, Yarnold PR, Kwaan HC, Masi DD, Sarode R, Raife TJ, Kiss JE, Raisch DW, Davidson C, Sadler JE, Ortel TL, Zheng XL, Kato S, Matsumoto M, Uemura M, Fujimura Y (2007). Two mechanistic pathways for thienopyridine-associated thrombotic thrombocytopenic purpura: A report from the Surveillance, Epidemiology, and Risk Factors for Thrombotic Thrombocytopenic Purpura (SERF-TTP) Research Group and the Research on Adverse Drug Events and Reports (RADAR) Project. *Journal of American College of Cardiology, 50*, 1138-1143.

[33]Soltysik RC, Yarnold PR (2010). Two-group MultiODA: Mixed-integer linear programming solution with bounded *M*. *Optimal Data Analysis, 1*, 31-37.

[34]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis, 6*, 26-42.

[35]Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement, 56*, 656-667.

[36]Lyons JS (1997). The evolving role of outcomes in managed health care. *Journal of Child and Family Studies, 6*, 1-8.

[37]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160.

[38]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice, 22*, 839-847.

[39]Donenberg GR, Bryant FB, Emerson E, Wilson HW, Pasch KE (2003). Tracing the roots of early sexual debut among adolescents in psychiatric care. *Journal of the American Academy of Psychiatry, 42*, 594-608.

[40]Kanter, AS, Spencer DC, Steinberg MH, Soltysik RC, Yarnold PR, Graham NM (1999). Supplemental vitamin B and progression to AIDS and death in black South African patients infected with HIV. *Journal of Acquired Immune Deficiency Syndrome, 21*, 252-257.

[41]Arozullah AM, Yarnold PR, Weinstein RA, Nwadiaro N, McIlraith TB, Chmiel JS, Sipler AM, Chan C, Goetz MB, Schwartz D, Bennett CL (2000). A new preadmission staging system for predicting in-patient mortality from HIV-associated *Pneumocystis carinii* pneumonia in the early-HAART era. *American Journal of Respiratory and Critical Care Medicine, 161*, 1081-1086.

[42]Coakley RM, Holmbeck GN, Bryant FB (2006). Constructing a prospective model of psychosocial adaptation in young adolescents with spina bifida: An application of optimal data analysis. *Journal of Pediatric Psychology, 31*, 1084-1099.

2155-0182/10/$3.00

[43]Taft CT, Pless AP, Stalans LJ, Koenen KC, King LA, King DW (2005). Risk factors for partner violence among a national sample of combat veterans. *Journal of Consulting and Clinical Psychology, 73*, 151-159.

[44]Green D, Hartwig D, Chen D, Soltysik RC, Yarnold PR (2003). Spinal cord injury risk assessment for thromboembolism (SPIRATE Study). *American Journal of Physical Medicine and Rehabilitation, 12*, 950-956.

[45]Stalans LJ, Yarnold PR, Seng M, Olson DE, Repp M. (2004). Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis. *Law & Human Behavior*, *28*, 53-271.

[46]Bryant FB, Yarnold PR (2014). Finding joy in the past, present, and future: The relationship between Type A behavior and savoring beliefs among college undergraduates. *Optimal Data Analysis*, *3*, 36-41.

[47]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, *22*, 875-885.

[48]Yarnold PR, Linden A. (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis*, *5*, 65-73.

[49]Cromley T, Lavigne JV (2008). Predictors and consequences of early gains in child psychotherapy. *Psychotherapy: Theory, Research, Practice, Training, 45*, 42-60.

[50]Snowden J, Leon S, Sieracki J (2008). Predictors of children in foster care being adopted: A classification tree analysis. *Children and Youth Services Review, 30*, 1318-1327.

[51]Belknap SM, Moore H, Lanzotti SA, Yarnold PR, Getz,M, Deitrick DL, Peterson A, Akeson J, Maurer T, Soltysik RC, Storm J (2008). Application of software design principles and debugging methods to an analgesia prescription reduces risk of severe injury from medical use of opioids. *Clinical Pharmacology and Therapeutics*, *84*, 385-392.

[52]Alshekhlee A, Ranawat N, Syed TU, Conway D, Ahmad SA, Zaiday OO (2010). National Institutes of Health Stroke Scale assists in predicting the need for percutaneous endoscopic gastrostomy tube placement in acute ischemic stroke. *Journal of Stroke and Cerebrovascular Diseases*, *19*, 347-352.

[53]Yarnold PR, Bennett CL (2016). Novometrics *vs*. multiple regression analysis: Age and clinical measures of PCP survivors, *Optimal Data Analysis*, *5*, 79-82.

[54]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, *23*, 703-712.

[55]Linden A, Yarnold PR (2017). Modeling time-to-event (survival) data using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1299-1308.

[56]Linden A, Yarnold PR (2018). Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice*, *24*, 380-387.

[57]Linden A, Yarnold PR (2018). Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *24*, 353-361.

Author Notes

No conflict of interest was reported.