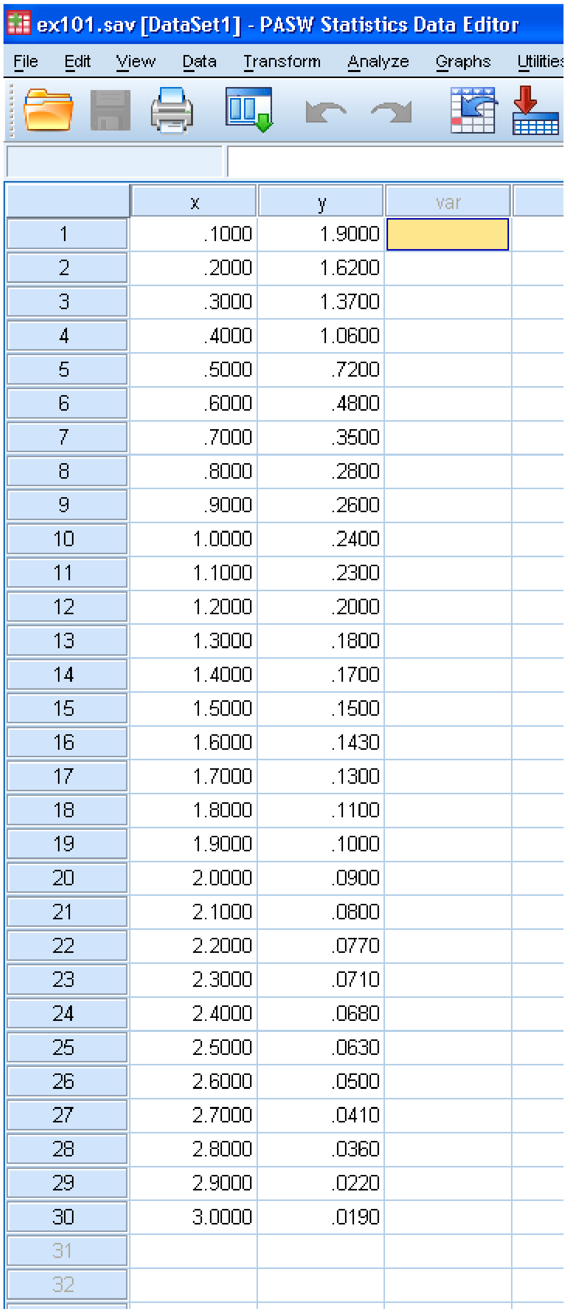
Model Specification Error…Are you straight, or do you have curves?
By Dr. Jon Starkweather, Research and Statistical Support consultant

The fact is; most of us have curves, even if we don’t want to acknowledge them. Model specification error generally refers to errors of omission and errors of inclusion, meaning; omitting crucial variables from the model or entering useless variables into the model. However, model specification error, or model misspecification, can also refer to the model *form* imposed on the data (e.g. *general linear* model). The purpose of this month’s article is to show how specification of different models to the same data can lead to meaningful differences in the interpretation of inferential analysis. The example provided is a very simple bivariate regression with an exaggerated pattern of data points. It is important to note that this example is simple and exaggerated specifically to illustrate how choice of model can make a difference in the results of an analysis. It is worth noting at the outset that with larger, more complex data which may contain less exaggerated patterns and indeed difficult to identify patterns, specification of the most appropriate model can become very important. These points are mentioned because when contemplating any type of model fitting analysis, one is expected to do a thorough job of exploring one’s data to discover the underlying relationships between variables of interest. During the process of initial data analysis one will likely discover the underlying pattern(s) in the data and proceed with the appropriate type of model. All of what follows can be duplicated in PASW/SPSS 18.



	x	y	var
1	.1000	1.9000	
2	.2000	1.6200	
3	.3000	1.3700	
4	.4000	1.0600	
5	.5000	.7200	
6	.6000	.4800	
7	.7000	.3500	
8	.8000	.2800	
9	.9000	.2600	
10	1.0000	.2400	
11	1.1000	.2300	
12	1.2000	.2000	
13	1.3000	.1800	
14	1.4000	.1700	
15	1.5000	.1500	
16	1.6000	.1430	
17	1.7000	.1300	
18	1.8000	.1100	
19	1.9000	.1000	
20	2.0000	.0900	
21	2.1000	.0800	
22	2.2000	.0770	
23	2.3000	.0710	
24	2.4000	.0680	
25	2.5000	.0630	
26	2.6000	.0500	
27	2.7000	.0410	
28	2.8000	.0360	
29	2.9000	.0220	
30	3.0000	.0190	
31			
32			

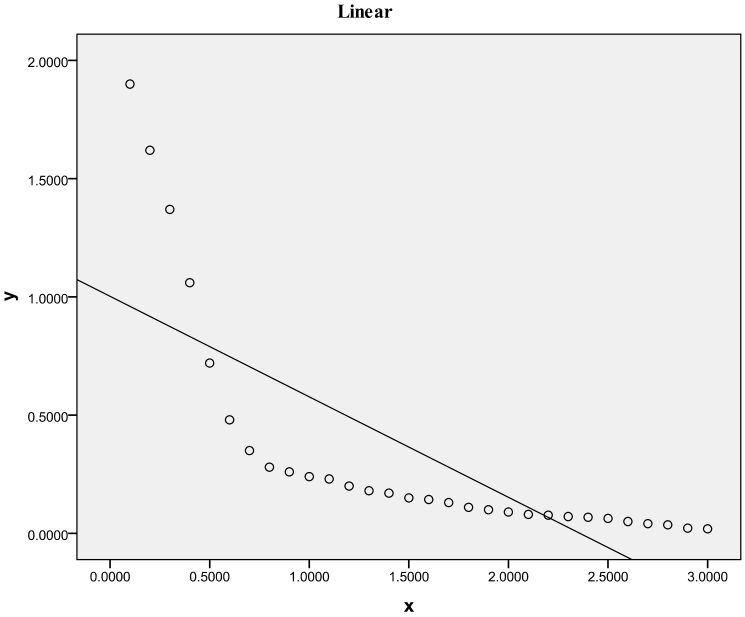
The current example utilizes two variables (x & y) each containing 30 data points (n = 30). The x variable is our predictor and the y variable is our outcome. If we apply a common **linear** regression

$$(1) \qquad y = bx + a$$

analysis to our data, as is often the default; we find a strong negative correlation (r = -.759, p < .001). The model summary table would show a moderate effect size or amount of variance accounted for (Adj. R^2 = .561) and our ubiquitous ANOVA table would indicate that this model’s R^2 is significantly different from zero; or stated another way, this model is better than simply using the mean value of x to predict new y scores, *F*(1, 28) = 38.043, *p* < .001. Remember, linear regression represents a *straight* line; it can increase, decrease, or remain flat—which would indicate no relationship between the two variables. At this point, we might be inclined to call it a day and be sufficiently satisfied with ourselves and our analysis. We could say our model does a fairly good job and provides us with a decent effect size (R^2) which indicates that we could predict with reasonable accuracy using our **linear** model:

(2) $$y = -.425x + 1.002$$

However, this would be precisely the pitfall this article is designed to illuminate. As we will see, there are a few other models that better characterize this data. For instance, if we simply view a scatter plot with our **linear** best fit regression line, we see clearly there may be other models more appropriate for this data.



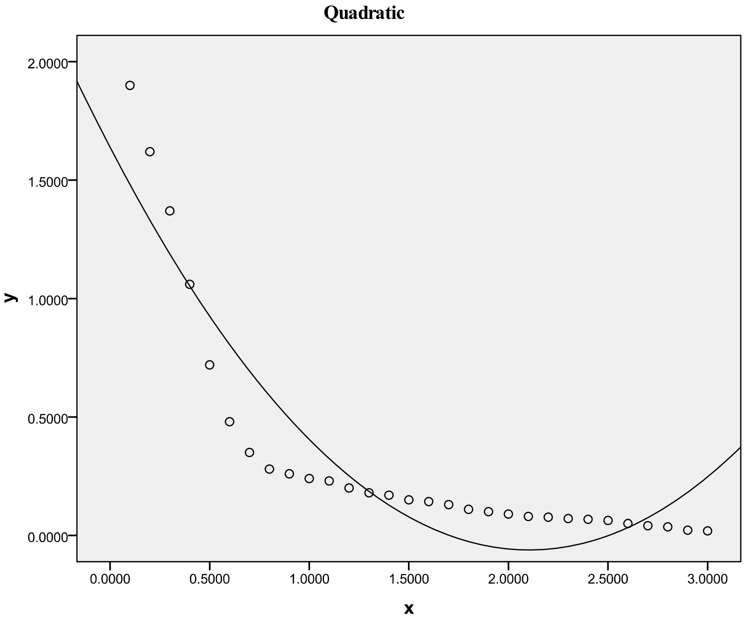
So, you may ask yourself "how do I better characterize the data?" And, after looking at the scatter plot above, you may tell yourself "a **quadratic** model would fit this data better." If we apply a **quadratic** regression

(3) $$y = b_1x^2 + b_2x + a$$

to our data, we find a distinct and meaningful increase in our effect size (Adj.R^2 = .847) as well as an increase in our *F* value from the ANOVA table, *F*(2, 27) = 81.397, *p* < .001; in fact the *F* value more than doubled. A quadratic regression represents a parabola, which can increase then decrease or decrease then increase. So, let's take a look at the scatter plot again, this time with a line representing our **quadratic** equation

(4) $$y = .384x^2 + -1.617x + 1.638$$

overlaying our data points.



Once again, we might now raise our chin and proclaim we have done a good job of modeling our data. After all, we've seen a substantial increase in our R^2, our F value, and we can see in this scatter plot that our model (represented as the line) better fits the data points.

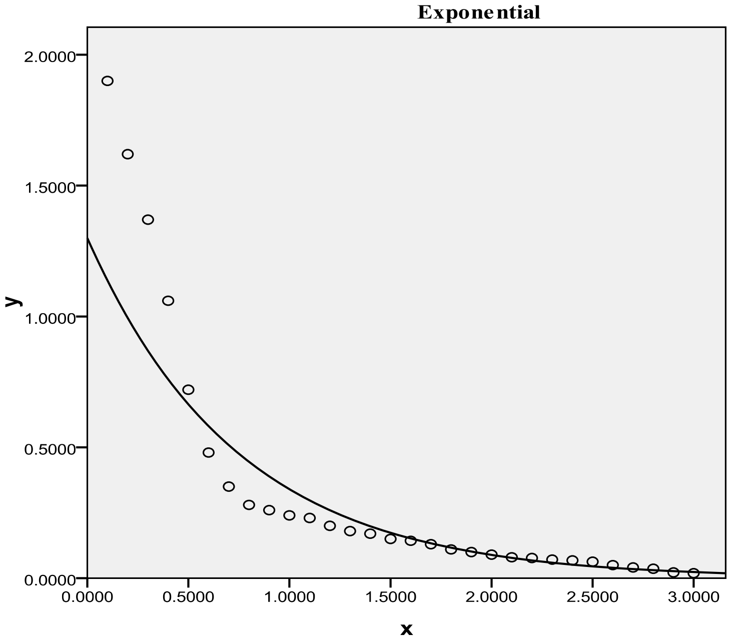
However, being the conscientious folks we are as data analysts and after having seen the differences we have thus far; we might be curious to see if we can find another model that better fits our data. So, if we apply an **exponential** regression

$$y = b^{x} + a \qquad (5)$$

to our data, we find a further improvement in our effect size (Adj. $R^2 = .948$). And with the **exponential** regression, we see our F value growing to enormous proportions, $F(1, 28) = 525.869$, $p < .001$. An **exponential** regression uses the predictor as an exponent and presents a line that can steeply increase or steeply decrease. So, let's take another look at the scatter plot; this time with our **exponential** equation

$$y = -1.341^{x} + 1.300 \qquad (6)$$

best fit line overlaying our data points.



Finally, we can stand up and speak with confidence that we have found an appropriate model for our data which accounts for 95% of the variance and fits our data very well. However, we may still be able to improve upon this with the application of yet another model.

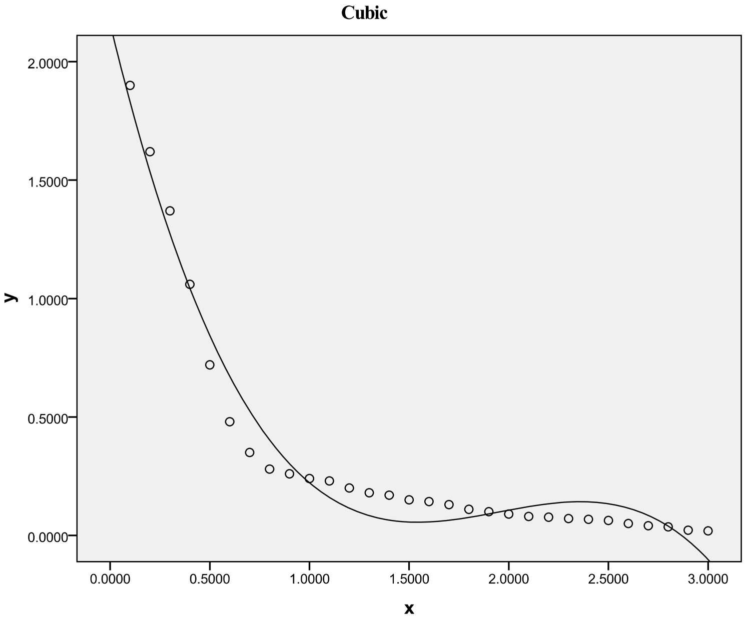
So, if we apply a **cubic** regression

(7) $$y = b_1x^3 + b_2x^2 + b_3x + a$$

to our data, we find a slightly higher effect size (Adj. $R^2 = .965$) and a slightly smaller (but still massive) F value, $F(3, 26) = 269.732$, $p < .001$. A **cubic** regression can steeply increase, steeply decrease, increase then decrease, or decrease then increase. If we use our **cubic** regression line

(8) $$y = -.319x^3 + 1.868x^2 + -3.486x + 2.160$$

to graphically display our model's fit on the data in a scatter plot, then we can see it fits slightly better than the previous model.



It appears as though we have squeezed every ounce of R^2 from our data as possible. However, there are other types of models; even for regression in PASW/SPSS—such as Logarithmic, Inverse, Power, Compound, S, Logistic, and Growth. For a comparison of each; one can utilize the 'Curve Estimation…' function in PASW/SPSS by clicking on → Analyze → Regression → Curve Estimation… just remember it may be beneficial to click on the 'Display ANOVA table' box in the 'Curve Estimation' dialog box. Clicking that box will show the model summary table, ANOVA summary table and coefficients table for each type of model being compared. Without checking that box, one gets a global 'Model Summary and Parameter Estimates' table such as this:

Model Summary and Parameter Estimates

Dependent Variable:y

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.576	38.043	1	28	.000	1.002	-.425		
Quadratic	.858	81.397	2	27	.000	1.638	-1.617	.384	
Cubic	.969	269.732	3	26	.000	2.160	-3.486	1.868	-.319
Exponential	.949	525.869	1	28	.000	1.300	-1.341		

The independent variable is x.

This global table can be a bit confusing if one compares how PASW designates each of the coefficients in comparison to how they are notated in the equations above. Notice in particular the coefficients in the table for the cubic model; where the table lists b1 and b3 which correspond to the third and first coefficients from left to right in the equation. This is why it is recommended to always check the box to display the ANOVA table; which also displays a more intuitive coefficients table for each model being compared—as the example below shows for the cubic model.

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.984	.969	.965	.092

The independent variable is x.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	6.824	3	2.275	269.732	.000
Residual	.219	26	.008		
Total	7.043	29			

The independent variable is x.

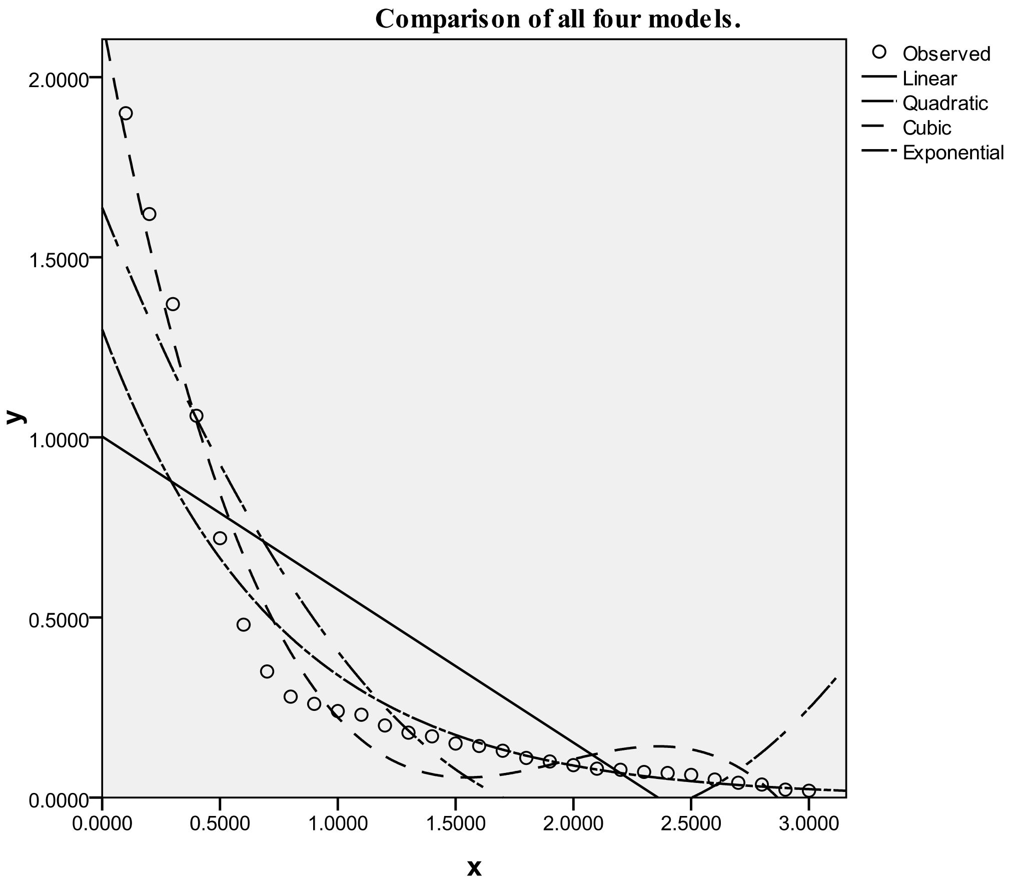
Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
x	-3.486	.210	-6.228	-16.606	.000
x ** 2	1.868	.156	10.661	11.973	.000
x ** 3	-.319	.033	-5.298	-9.634	.000
(Constant)	2.160	.076		28.271	.000

Notice with this style of coefficients table the coefficients are designated not with b1, b2, b3; but rather with the exponents, such as x, x**2, x**3, which clarifies where each coefficient should go in the cubic regression equation. Also, the model summary table for each model provides not just R^2, but also Adj. R^2.

Please note we have been using Adj. R^2 throughout this article as a metric for comparing our models because it is readily available, easy to interpret, and is extremely well known. Most are familiar with the shrinkage which makes Adj. R^2 preferable over R^2. However, when comparing models, Adj. R^2 is not much better than the base R^2. One rule of thumb is, if there is a .10 or 10% difference between Adj. R^2 and R^2 then overfitting is a concern (Harrell, Lee, & Mark, 1996). Therefore; it is recommended the Akaike's information criterion (AIC; Akaike, 1974) or the Bayesian information criterion (BIC; Schwarz, 1978) be used instead—both of which are much more appropriate for assessing a model's worth and comparing multiple model's fit (Kass & Raftery, 1995).

Although a bit confusing, the following scatter plot was produced using the 'Curve Estimation…' function and reflects each of the four models reviewed here.



Please also note that this article does not intend to represent the complete range of techniques available for extracting the maximum information from a set of data or a regression analysis approach to data. There are other types of regression analysis and techniques available that may allow the researcher to extract a more complete picture of the phenomena of interest from the data. Regression analysis examples include but are not limited to; Tobit, Quantile, Partial Least Square, Binary Logistic, Multinomial Logistic, Ordinal, Probit, 2-Stage Least Square, as well as re-sampling techniques for reducing bias such as bootstrapping. If there is one message this author hopes the reader will take from this article, it is this; do not fall into the trap of complacency and rely exclusively on the default settings or analysis of your software, be thorough and un-intimidated by the plethora of non-traditional data analytic techniques at your disposal.

Until next time, remember; this land is your land, this land is my land… and I'll be at Alice's Restaurant.

References

Akaike, H. (1974). A new look at the statistical model identification. *I.E.E.E. Transactions on automatic control, AC 19,* 716 – 723. (1) (2) (3)

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*, 361 – 387. (1) (2) (3)

Kass, R. E., & Raftery, A., E. (1995). Bayses factors. *Journal of the American Statistical Association*, *90*, 773 – 795. (1) (2) (3)

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461 – 464. (1) (2) (3)