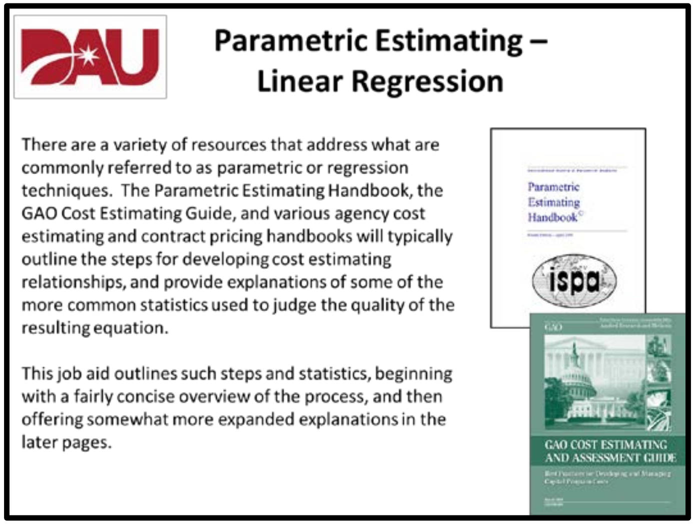
Parametric Estimating – Nonlinear Regression

The term “nonlinear” regression, in the context of this job aid, is used to describe the application of linear regression in fitting nonlinear patterns in the data. The techniques outlined here are offered as samples of the types of approaches used to fit patterns that some might refer to as being “curvilinear” in nature.

This job aid is intended as a complement to the Linear Regression job aid which outlines the process of developing a cost estimating relationship (CER), addresses some of the common goodness of fit statistics, and provides an introduction to some of the issues concerning outliers. The first 6 steps from that job aid are cited on the next page for reference.



Parametric Estimating – Linear Regression

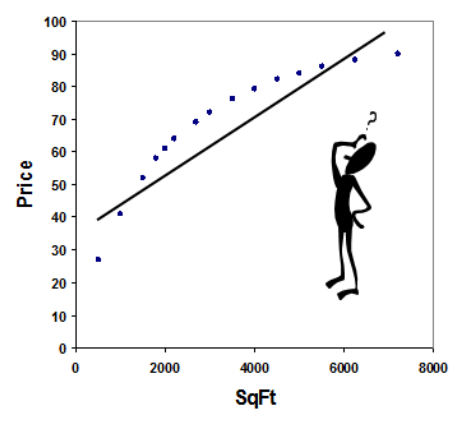
There are a variety of resources that address what are commonly referred to as parametric or regression techniques. The Parametric Estimating Handbook, the GAO Cost Estimating Guide, and various agency cost estimating and contract pricing handbooks will typically outline the steps for developing cost estimating relationships, and provide explanations of some of the more common statistics used to judge the quality of the resulting equation.

This job aid outlines such steps and statistics, beginning with a fairly concise overview of the process, and then offering somewhat more expanded explanations in the later pages.

Parametric Estimating Handbook
ispa
GAO COST ESTIMATING AND ASSESSMENT GUIDE

Nonlinear Regression

Nonlinear Regression

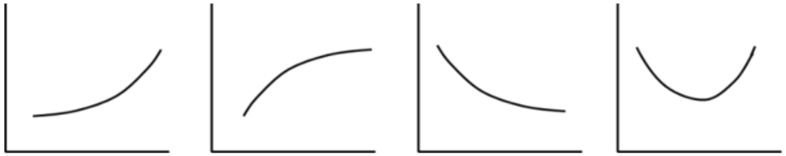


1. **Identification of Cost Drivers**
2. **Specification**
3. **Data Collection**
4. **Normalization**
5. **Graphical Analysis**
6. **Fitting the Data (Nonlinear)**
   a) Transformation on X
   b) Quadratic Equation
   c) Power Equation

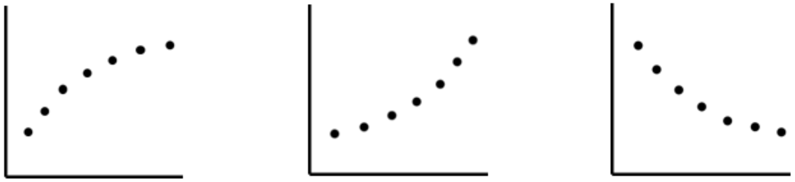
These approaches are sometimes called “intrinsically linear” in that the data is transformed or modeled using linear relationships. Some “nonlinear” trends are “not linear” in that the data can’t be transformed or modeled with a linear function (e.g. step functions).

When would we consider a Nonlinear approach?

1) The expectation or specification by subject matter expert

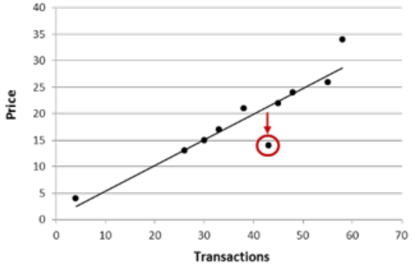


2) Observation based on graphical analysis of the data

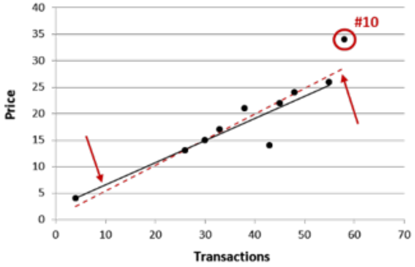


When would we consider? 3) As a remedy for a:

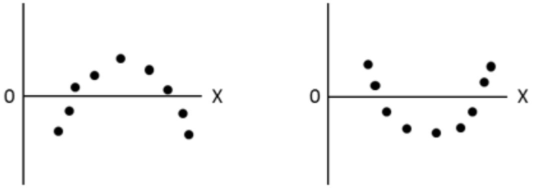
Prediction Problem



Influential Observation



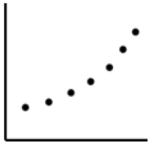
Nonlinearity in the Residual Plots



Nonlinear Techniques

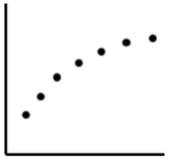
X Transformations

Increasing at an increasing rate



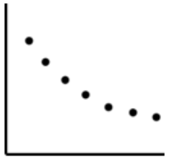
X^2

Increasing at a decreasing rate



$\sqrt{X}$

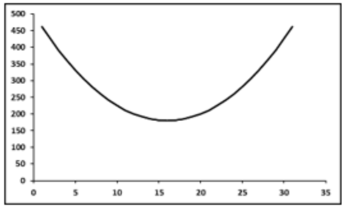
Decreasing at a decreasing rate



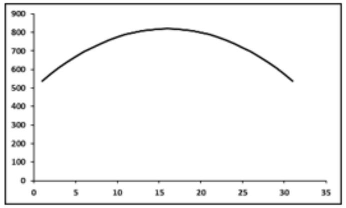
$\frac{1}{X}$

Quadratic Equation

$$\hat{Y} = b_0 + b_1X_1 + b_2X_1^2$$



…this when b_2 is **positive**…



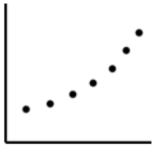
…and this when b_2 is **negative**

Nonlinear Regression (continued)

Nonlinear Techniques

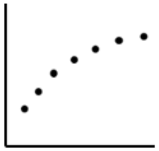
The Power Model $\widehat{Y} = b_0(X)^{b_1}$

Increasing at an increasing rate



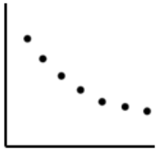
$b_1 > 1$

Increasing at a decreasing rate



$0 < b_1 < 1$

Decreasing at a decreasing rate



$b_1 < 0$

Power Model a.k.a. Log-Log Model

X and Y transformation using either Log X, Log Y or using LN X, LN Y

Common Logarithm (LOG)
- uses base 10
- the LOG of a number is the power to which 10 must be raised to obtain that number

$\text{LOG}100 = 2$ (i.e. $10^2 = 100$)

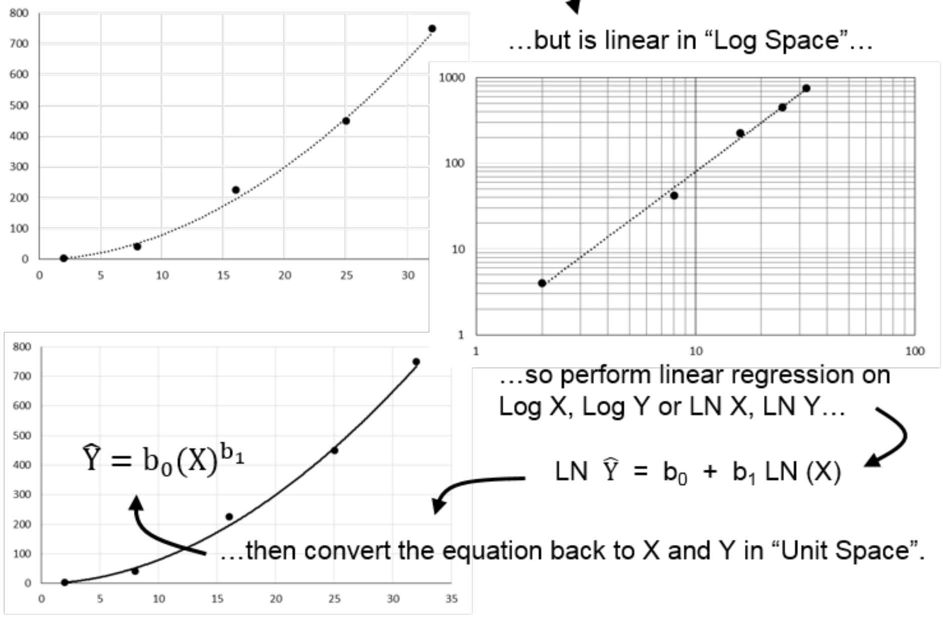
$\text{LOG}1000 = 3$ (i.e. $10^3 = 1000$)

Natural Logarithm (LN)
- uses base e (2.71828…)
- the LN of a number is the power to which "e" must be raised to obtain that number

$\text{LN}100 = 4.605$ (i.e. $2.71828^{4.605} = 100$)

$\text{LN}1000 = 6.9078$ (i.e. $2.71828^{6.9078} = 1000$)

Data is nonlinear in "Unit Space"…



…but is linear in "Log Space"…

…so perform linear regression on Log X, Log Y or LN X, LN Y…

$\text{LN } \widehat{Y} = b_0 + b_1 \text{ LN } (X)$

…then convert the equation back to X and Y in "Unit Space".

Creating the Power Model – Regress Log X, Log Y or LN X, LN Y such as:

Y	X	LN(Y)	LN (X)
4	2	1.3863	0.6931
42	8	3.7377	2.0794
225	16	5.4161	2.7726
450	25	6.1092	3.2189
750	32	6.6201	3.4657

Equation in "**Log**" Space

$\text{LN } \widehat{Y} = -0.0102 + 1.9069 \text{ LN } (X)$

Converting an equation from Log Space to Unit Space

Take the antilog of the intercept ($e^{-0.0102} = 0.9899$)

Slope in log space 1.9069 LN(X) becomes the exponent in unit space $X^{1.9069}$

Equation in "**Unit**" space (i.e. Power Model): $\widehat{Y} = 0.9899\ (X)^{1.9069}$

The Standard Error (SE) in "Unit Space" for the Power Model

(X) Independent	(Y) Dependent	$\widehat{Y}$ Predicted	$Y - \widehat{Y}$ Residual	$(Y - \widehat{Y})^2$ Residual^2
2	4	3.71	0.29	0.08
8	42	52.20	-10.20	104.04
16	225	195.76	29.24	854.98
25	450	458.48	-8.48	71.91
32	750	734.11	15.89	252.49
			26.74	1283.50

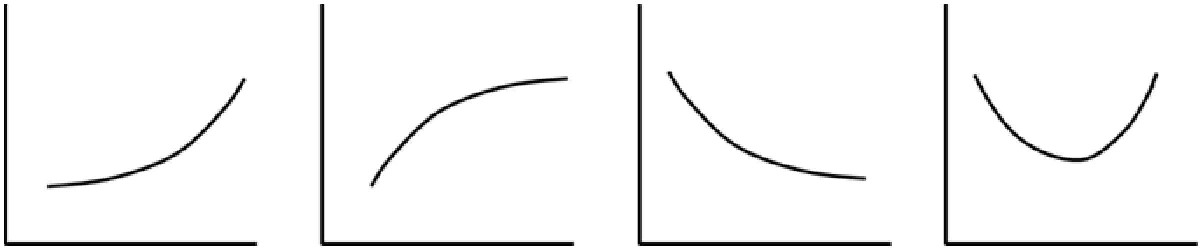
$$\text{Variance (MSE)} = \frac{\text{SSE}}{\text{DF}} = \frac{\sum(Y - \widehat{Y})^2}{n - 2} = \frac{1283.50}{3} = 427.83$$

$$\text{Standard Error (SE)in } \textbf{Unit Space} = \sqrt{\text{Variance}} = \sqrt{427.83} = 20.68$$

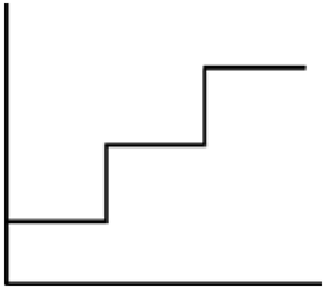
$$\overline{Y} = 294.20 \qquad \text{CV} = \frac{\text{SE}}{\overline{Y}} = \frac{20.68}{294.20} = .0703 \text{ or } 7.03\%$$

What do we mean by the term “nonlinear”?

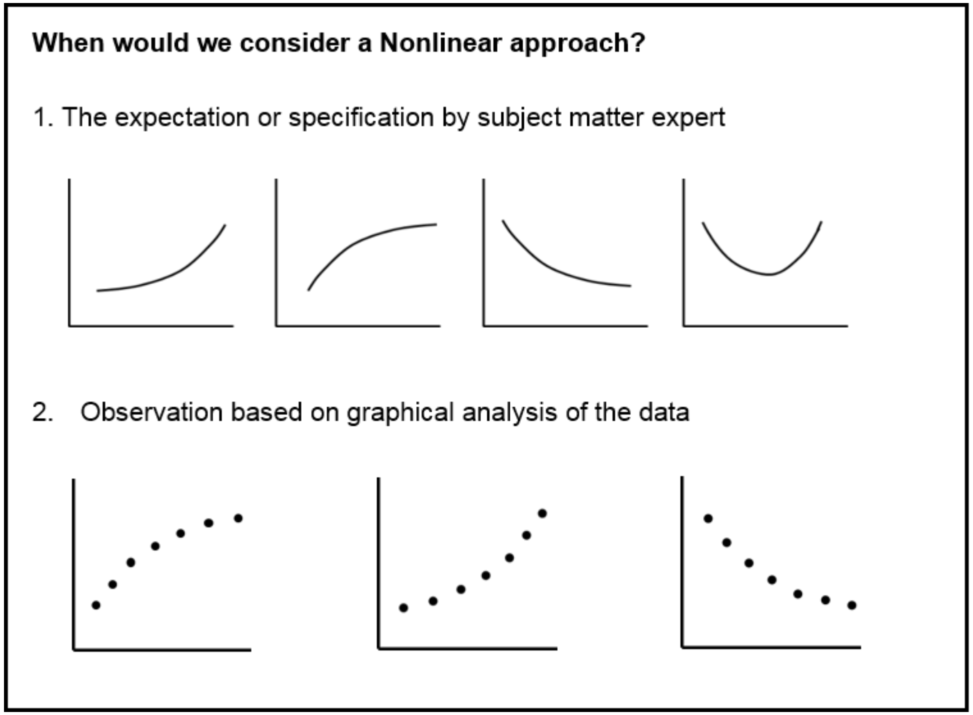
This job aid will address several techniques intended to fit patterns, such as the ones immediately below, that will be described here as being nonlinear or curvilinear (i.e. consisting of a curved line). These types of shapes are sometimes referred to as being “intrinsically linear” in that they can be “linearized” and then fit with linear equations.



For our purposes we will describe the shape below as being “not linear”. The techniques described here cannot be used to fit these types of relationships.



When would we consider a Nonlinear approach?



When would we consider a Nonlinear approach?

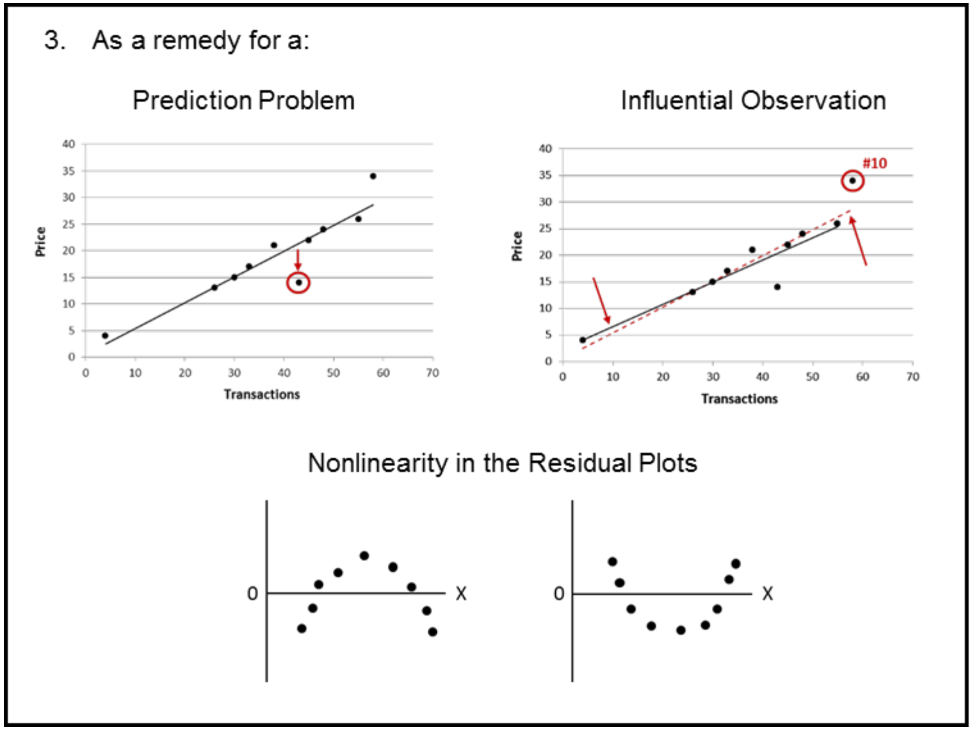
1. The expectation or specification by subject matter expert

2. Observation based on graphical analysis of the data

The Linear Regression job aid suggests that the first step in developing a cost estimating relationship would be to involve your subject matter experts in the identification of potential explanatory (X) variables. The second step would be to specify what the expected relationships would look like between the dependent (Y) variable and potential X variables. Those expectations may identify the need for a nonlinear technique.

It’s also a good practice to scatterplot the data and observe whether the data is consistent with expectations; or, if lacking specific expectations, whether the data itself makes a compelling case to consider either a linear technique or nonlinear technique.

Other reasons to consider a Nonlinear approach



3. As a remedy for a:

Prediction Problem

Influential Observation

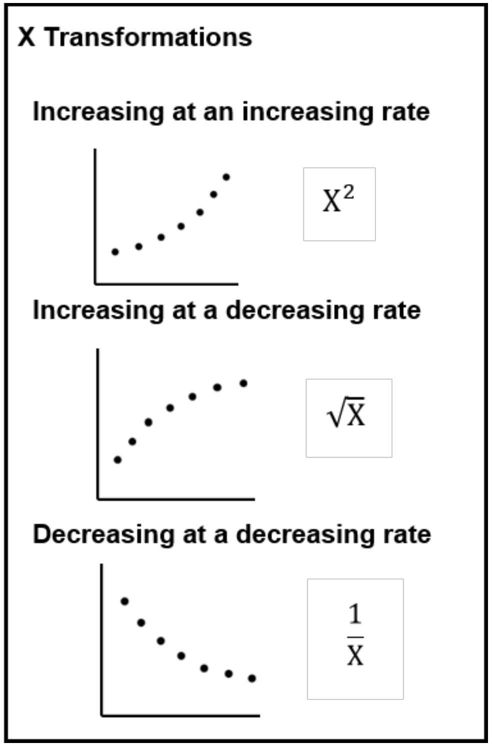
Nonlinearity in the Residual Plots

The Linear Regression job aid identifies some of the potential problems that you might experience with an equation such as: a data point that is more poorly predicted by the equation that the other data points; an influential observation; and residuals evidencing a pattern that would suggest nonlinearity in the data.

There were a number of investigative steps suggested with each of these types of problems, one of those steps would have you consider the possibility that the data had not been properly fit (e.g. a linear equation had been used to fit data that was predominately nonlinear in nature) in which case a nonlinear fitting technique might be appropriate.

Fitting Data using an X Transformation

The term “transformation” is used in this job aid to describe the mathematical operations that can be performed on an X variable, Y variable, or X and Y variable such that an otherwise existing nonlinear relationship between X and Y can be made more linear by virtue of the transformation. A linear regression is then performed using the transformed variables. The illustrations below deal with transforming only the X variable.



X Transformations

Increasing at an increasing rate

X^2

Increasing at a decreasing rate

$\sqrt{X}$

Decreasing at a decreasing rate

$\frac{1}{X}$

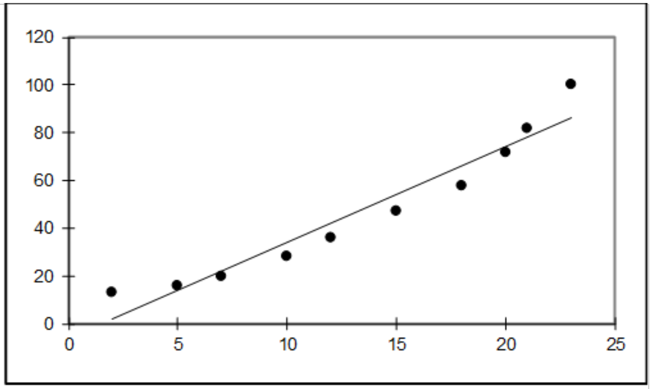
The first example shows a pattern between X and Y that we will call “increasing at an increasing rate”. One possible approach in fitting this data would be to do a linear regression with Y and X squared.

The second case shows a pattern between X and Y that could be called “increasing at an decreasing rate”. One technique would be to fit this data by regressing Y against the square root of X.

The third example is a pattern between X and Y we might call “decreasing at an decreasing rate”. In this case, regressing Y against the reciprocal of X might result in a better fit.

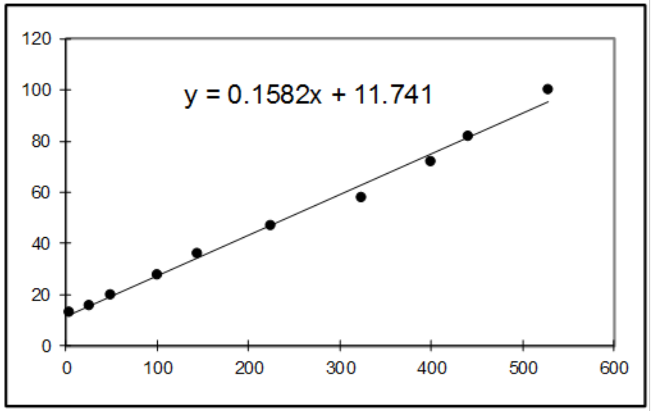
Using an X Transformation

X	Y
2	13
5	16
7	20
10	28
12	36
15	47
18	58
20	72
21	82
23	100



The relationship between X and Y appears to be increasing at an increasing rate. This would suggest an X squared transformation.

X	Y
4	13
25	16
49	20
100	28
144	36
225	47
324	58
400	72
441	82
529	100



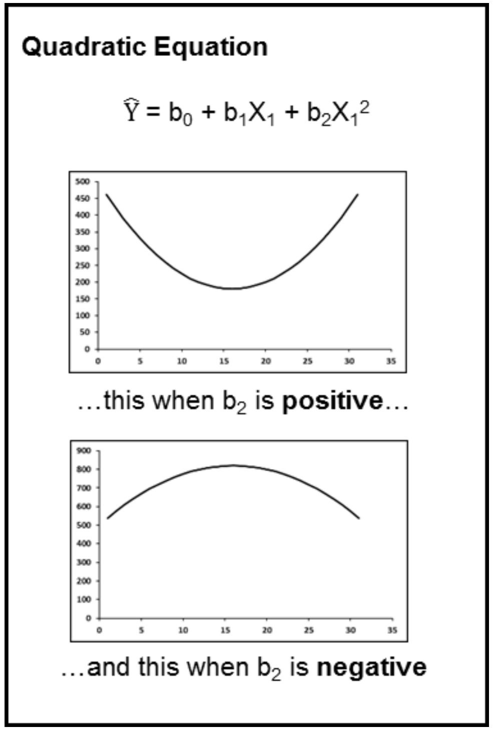
Notice that only the X values have been squared, the Y values remain the same. The result is a more linear relationship which can now be better fit with linear regression.

It's important to note that regardless of the application you might use, the application cannot distinguish that the values you are fitting are X squared and not X.

In applying the equation you must substitute X squared (in this case) for X, or whatever transformed X was used in creating the equation.

Fitting Data using a Quadratic Equation

The quadratic equation is a linear regression where the same X variable is used twice, once in it's untransformed state, and second as the square of that X variable



Quadratic Equation

$$\hat{Y} = b_0 + b_1X_1 + b_2X_1^2$$

…this when b_2 is **positive**…

…and this when b_2 is **negative**

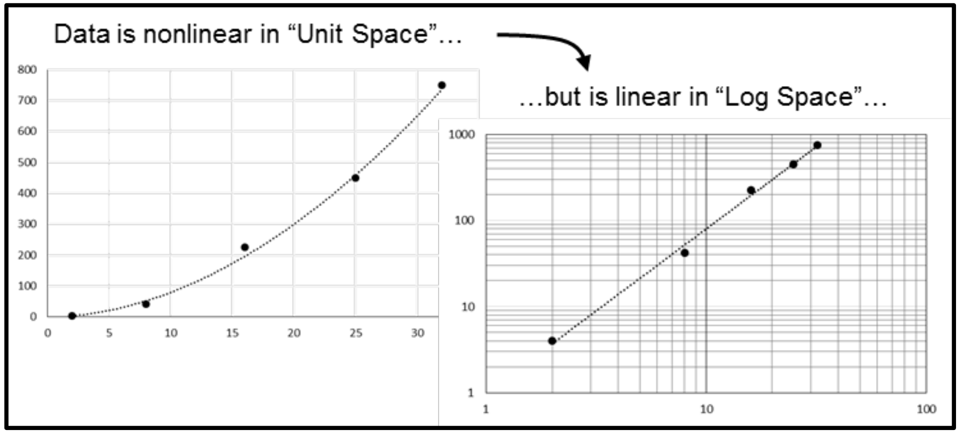
The equation produces the "right-side up" parabola when the coefficient on X squared is positive, and it produces the "upside down" parabola when the coefficient on X squared is negative.

If you were to bisect each of the two parabolas you would note that the quadratic can fit the previously mentioned "decreasing at a decreasing rate", "increasing at an increasing rate", and "increasing at a decreasing rate" patterns within certain ranges of the equation. Since the patterns are in fact range dependent in the quadratic equation, it's particularly important not to extrapolate beyond the range of the data, otherwise unexpected results would occur.

Since the same X variable is being used twice in the equation, it is inevitable that correlation will exist between X and X squared. Although the correlation exists, it does not pose some of the issues as when the correlation is between different X variables. For more on correlation between the X variables, and equations with multiple X variables, see the Multiple Regression job aid.

The Power Model $\widehat{Y}_X = b_0(X)^{b_1}$

It’s been observed that where a nonlinear pattern exists between the X and Y variables, the pattern between Log X and Log Y tends to be much more linear. The Power model is the result of a logarithmic transformation of both the X and Y variables.


Data is nonlinear in “Unit Space”…

…but is linear in “Log Space”…

The transformation on X and Y can be done using either the common (base 10) logarithm (LOG) or the natural (base e) logarithm (LN).

Since the regression is performed using either the LOG or LN values of X and Y, you may also see the power model referred to as the log-linear model or the log-log model.

Note, the graph on the upper left is referred to as Cartesian space, where the values of the variables exist in their normal “units” of measure (e.g. dollars, hours, pounds, horsepower). We will call this “Unit” space, in contrast to the logarithmic scale on the upper right which we will call “Log” space.

Power Model a.k.a. Log-Log Model

X and Y transformation using either Log X, Log Y or using LN X, LN Y

Common Logarithm (LOG)
- uses base 10
- the LOG of a number is the power to which 10 must be raised to obtain that number

$$\text{LOG}\,100 = 2 \ \ (\text{i.e.}\ 10^2 = 100)$$

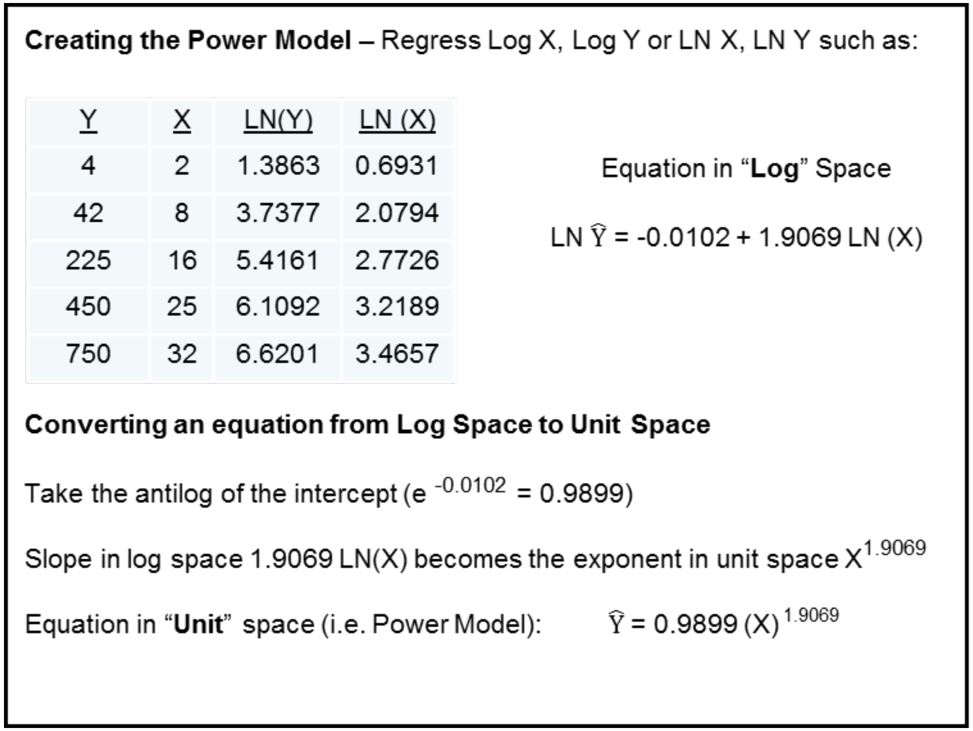
$$\text{LOG}\,1000 = 3 \ \ (\text{i.e.}\ 10^3 = 1000)$$

Natural Logarithm (LN)
- uses base e (2.71828…)
- the LN of a number is the power to which “e” must be raised to obtain that number

$$\text{LN}\,100 = 4.605 \ \ (\text{i.e.}\ 2.71828^{4.605} = 100)$$

$$\text{LN}\,1000 = 6.9078 \ \ (\text{i.e.}\ 2.71828^{6.9078} = 1000)$$

Creating the Power Model



Creating the Power Model – Regress Log X, Log Y or LN X, LN Y such as:

Y	X	LN(Y)	LN (X)
4	2	1.3863	0.6931
42	8	3.7377	2.0794
225	16	5.4161	2.7726
450	25	6.1092	3.2189
750	32	6.6201	3.4657

Equation in "**Log**" Space

$LN\ \hat{Y} = -0.0102 + 1.9069\ LN\ (X)$

Converting an equation from Log Space to Unit Space

Take the antilog of the intercept ($e^{-0.0102} = 0.9899$)

Slope in log space 1.9069 LN(X) becomes the exponent in unit space $X^{1.9069}$

Equation in "**Unit**" space (i.e. Power Model): $\hat{Y} = 0.9899\ (X)^{1.9069}$

In the example, linear regression is performed on the natural logarithm (LN) of X and Y. The resulting equation is linear in what we will call "log space", i.e. between LN(X) and LN(Y).

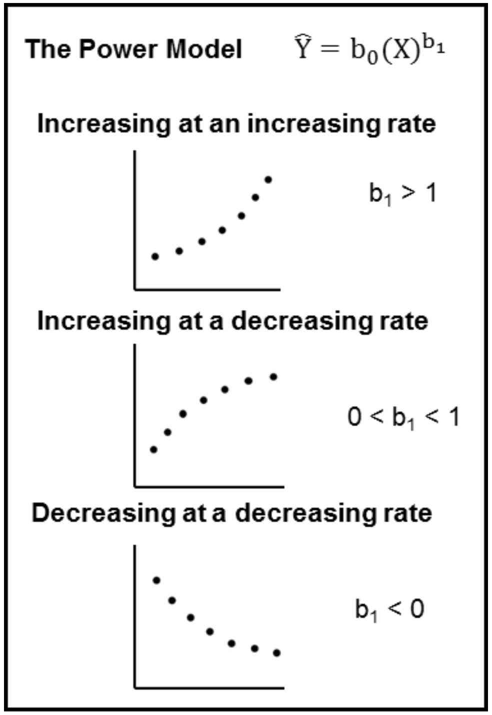
Since we began by taking the logs of X and Y, the process of converting back to X and Y will require us to take the antilog of the equation.

From the equation in log space we take the antilog of the intercept. The X variable's coefficient (slope) become the X variable's exponent. The value of Y now becomes the product of the terms, lending the equation to sometimes being called a multiplicative equation subject to a multiplicative error term.

In a linear equation the b_0 term represents the intercept, i.e. the value of Y when X is equal to zero (0.0). In the power model the b_0 term represents the value of Y when X is equal to one (1.0).

Exponents of the Power Model

The power equation is a popular convention when modeling nonlinear or curvilinear patterns due in part to the ability of the equation to produce three different curve shapes by simply varying the value of the exponent.



The first example shows a pattern between X and Y that we will call “increasing at an increasing rate”. An exponent greater than one (1.0) will produce these types of patterns.

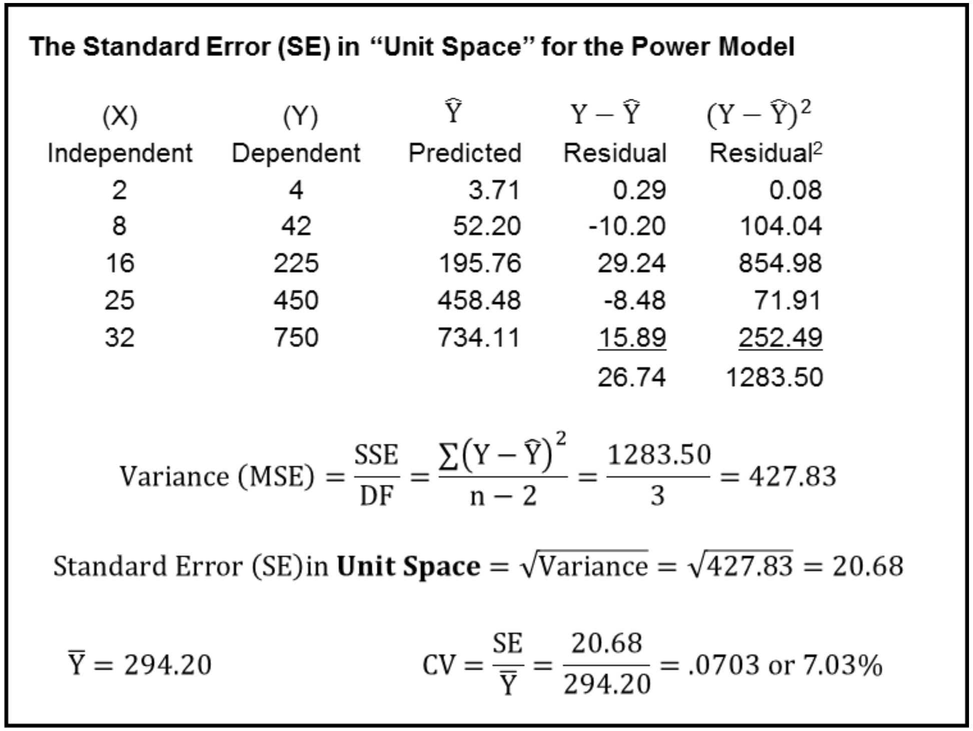
The second case shows a pattern between X and Y that could be called “increasing at an decreasing rate”. The equation will produce theses patterns if the value of the exponent is between zero (0.0) and one (1.0).

The third example is a pattern between X and Y we might call “decreasing at an decreasing rate”. An exponent less than zero (0.0), i.e. negative, will produce this shape.

The Standard Error for the Power Model

Since the linear regression was performed on LN(X) and LN(Y), the standard error reported on the regression output corresponds to the log linear equation. In order to put the standard error into context of Y rather than LN(Y), a standard error in "unit" space needs to be calculated.

The process follows the normal convention with a comparison of the actual value of Y and the predicted value of Y, in this case the predicted value being the result of entering the X values into the power model: $\hat{Y} = 0.9899\ (X)^{1.9069}$



The Standard Error (SE) in "Unit Space" for the Power Model

(X) Independent	(Y) Dependent	$\hat{Y}$ Predicted	$Y - \hat{Y}$ Residual	$(Y - \hat{Y})^2$ Residual^2
2	4	3.71	0.29	0.08
8	42	52.20	-10.20	104.04
16	225	195.76	29.24	854.98
25	450	458.48	-8.48	71.91
32	750	734.11	15.89	252.49
			26.74	1283.50

$$\text{Variance (MSE)} = \frac{\text{SSE}}{\text{DF}} = \frac{\sum(Y-\hat{Y})^2}{n-2} = \frac{1283.50}{3} = 427.83$$

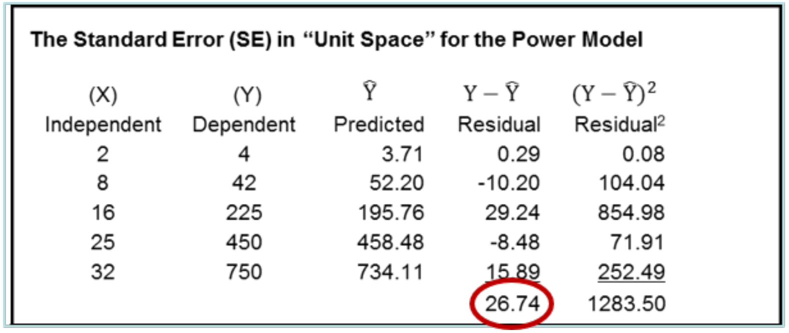
$$\text{Standard Error (SE)in } \textbf{Unit Space} = \sqrt{\text{Variance}} = \sqrt{427.83} = 20.68$$

$$\overline{Y} = 294.20 \qquad \text{CV} = \frac{\text{SE}}{\overline{Y}} = \frac{20.68}{294.20} = .0703 \text{ or } 7.03\%$$

If the Y variable was in $K for example, we could say the average estimating error is around $20.68K. We could then calculate the coefficient of variation (CV) and state that, relatively speaking, there is an average estimating error of around 7%.

(In both cases some latitude has been taken with the exact meaning of a standard error.)

The “Approximate” Standard Error for the Power Model

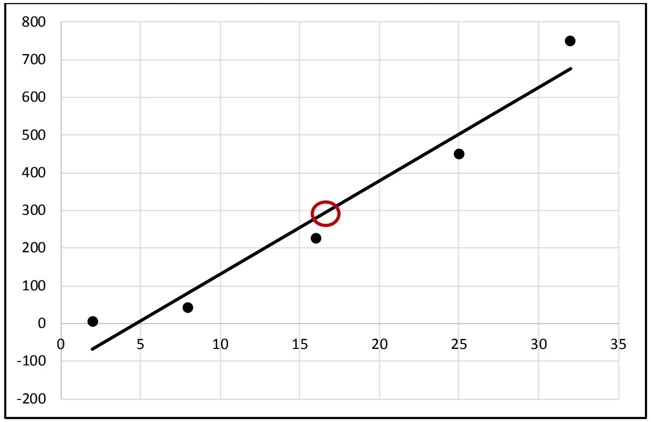


The Standard Error (SE) in “Unit Space” for the Power Model

(X)	(Y)	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
Independent	Dependent	Predicted	Residual	Residual²
2	4	3.71	0.29	0.08
8	42	52.20	-10.20	104.04
16	225	195.76	29.24	854.98
25	450	458.48	-8.48	71.91
32	750	734.11	15.89	252.49
			26.74	1283.50

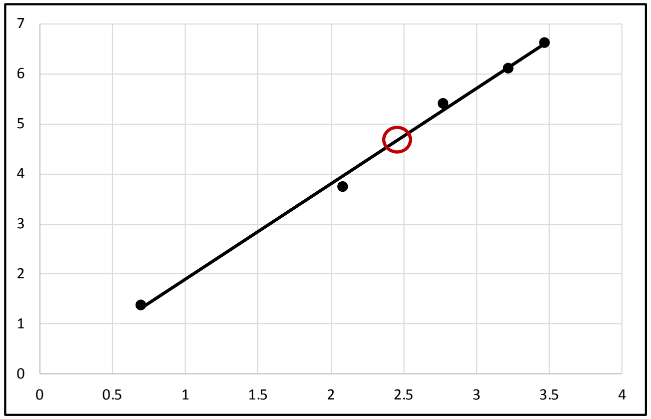
Notice that the residuals do not sum to zero as they would have in a linear equation.

	X	Y
	2	4
	8	42
	16	225
	25	450
	32	750
Mean	**16.6**	**294.2**



A linear equation is fitted through the means of X and Y.

	LN (X)	LN(Y)
	0.6931	1.3863
	2.0794	3.7377
	2.7726	5.4161
	3.2189	6.1092
	3.4657	6.6201
Mean	**2.4459**	**4.6539**



A log linear equation is fitted to the means of LN(X) and LN(Y).

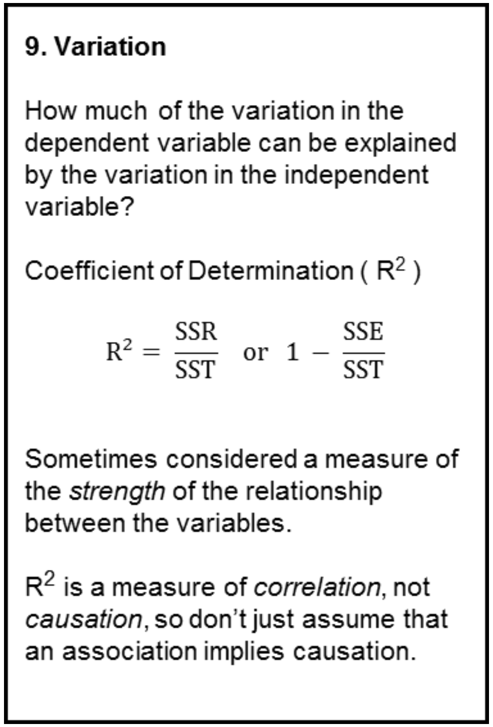
When the antilog is taken of the log linear equation to derive the power model, a “bias” occurs such that the residuals no longer sum to zero.

Consequently, the bias affects to some degree the accuracy of the unit space SE calculation.

The R Squared in Unit Space for the Power Model

The Linear Regression job aid (shown) notes that the R squared can be calculated by dividing the SSR (explained variation) by the SST (total variation). The resulting value is interpreted as the variation in Y explained by the variation in X.

Equivalently, in the linear equation, R squared can be calculated by taking one (1.0) minus the SSE (unexplained variation) divided by the SST. However, due to the bias previously noted with the conversion from log space to unit space, this equivalency no longer holds true when attempting to calculate the unit space R squared.



9. Variation

How much of the variation in the dependent variable can be explained by the variation in the independent variable?

Coefficient of Determination (R^2)

$$R^2 = \frac{\text{SSR}}{\text{SST}} \quad \text{or} \quad 1 - \frac{\text{SSE}}{\text{SST}}$$

Sometimes considered a measure of the *strength* of the relationship between the variables.

R^2 is a measure of *correlation*, not *causation*, so don't just assume that an association implies causation.

In the article "The Trouble with R^2" by Book and Young, they propose representing the R squared by taking the square of the Pearson R calculation (shown below).

$$r = \frac{\Sigma\left[(\mathrm{Y_i} - \overline{\mathrm{Y}})\left(\hat{\mathrm{Y}}_\mathrm{i} - \overline{\hat{\mathrm{Y}}}\right)\right]}{\sqrt{\Sigma(\mathrm{Y_i} - \overline{\mathrm{Y}})^2\,\Sigma\left(\hat{\mathrm{Y}}_\mathrm{i} - \overline{\hat{\mathrm{Y}}}\right)^2}}$$

The square of the Pearson R could be interpreted as the variation between the actual Y values and the predicted Y values that is explained by the equation.