Curve Fitting using Linear and Nonlinear Regression

By Jim Frost — 23 Comments

In regression analysis, curve fitting is the process of specifying the model that provic the best fit to the specific curves in your dataset. Curved relationships between variables are not as straightforward to fit and interpret as linear relationships.

For linear relationships, as you increase the independent variable by one unit, the m of the dependent variable always changes by a specific amount. This relationship ho true regardless of where you are in the observation space.

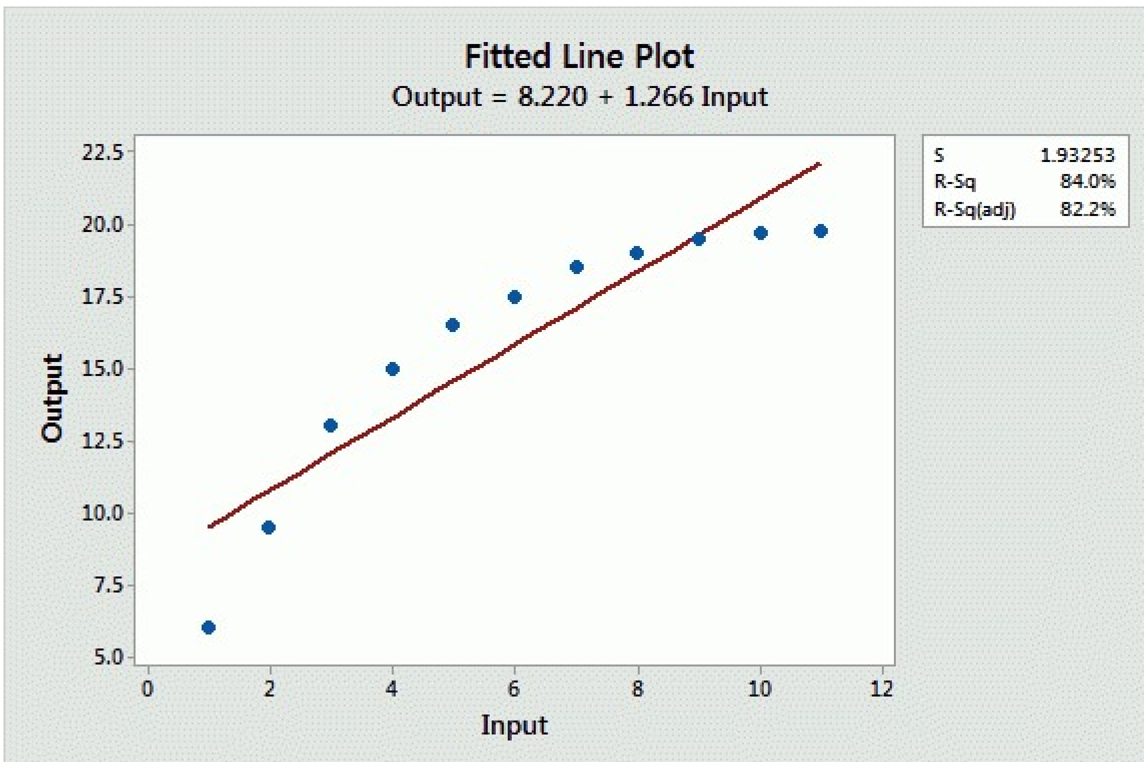
Unfortunately, the real world isn't always nice and neat like this. Sometimes your da have curved relationships between variables. In a curved relationship, the change in dependent variable associated with a one unit shift in the independent variable vari based on the location in the observation space. In other words, the effect of the independent variable is not a constant value.

Read my post where I discuss how to interpret regression coefficients for both linea and curvilinear relationships to see this in action.

In this post, I cover various curve fitting methods using both linear regression and nonlinear regression. I'll also show you how to determine which model provides the best fit.

Why You Need to Fit Curves in a Regression Model

The fitted line plot below illustrates the problem of using a linear relationship to fit a curved relationship. The R-squared is high, but the model is clearly inadequate. You need to do curve fitting!



When you have one independent variable, it's easy to see the curvature using a fitted line plot. However, with multiple regression, curved relationships are not always so apparent. For these cases, residual plots are a key indicator for whether your model adequately captures curved relationships.

If you see a pattern in the residual plots, your model doesn't provide an adequate fit the data. A common reason is that your model incorrectly models the curvature.

Plotting the residuals by each of your independent variables can help you locate the curved relationship.

Related post: Check Your Residual Plots to Ensure Trustworthy Results!

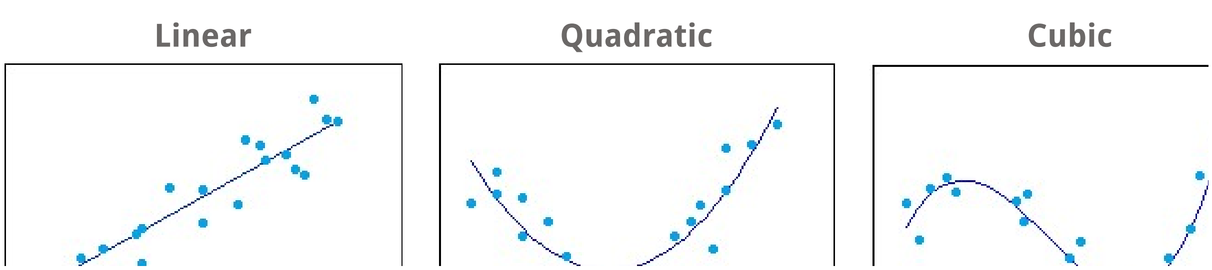
In others cases, you might need to depend on subject-area knowledge to do curve fitting. Previous experience or research can tell you that the effect of one variable or another varies based on the value of the independent variable. Perhaps there's a lim threshold, or point of diminishing returns where the relationship changes?

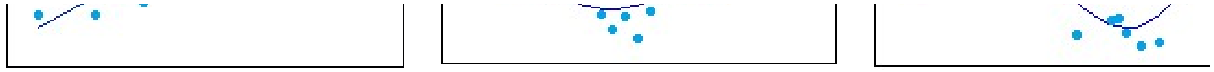
To compare curve fitting methods, I'll fit models to the curve in the fitted line plot ab because it is not an easy fit. Let's assume that these data are from a physical process with very precise measurements. We need to produce accurate predictions of the output for any specified input. You can download the CSV dataset for these example CurveFittingExample.

Curve Fitting using Polynomial Terms in Linear Regression

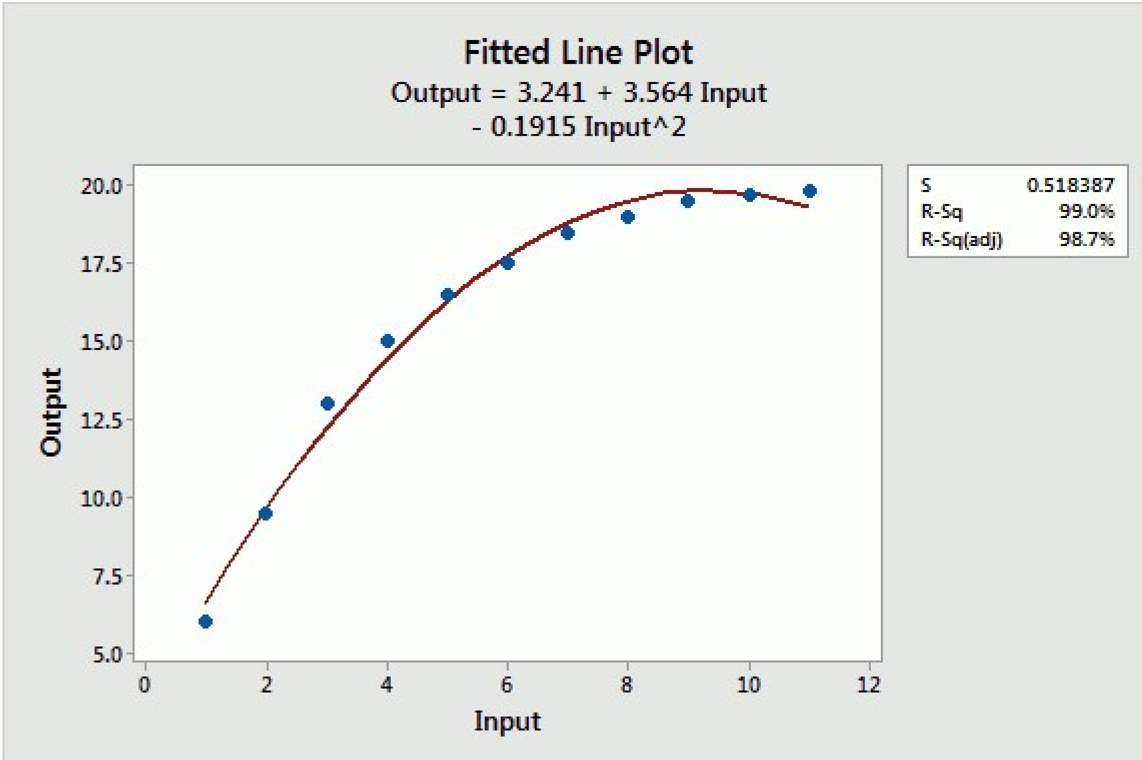
Despite its name, you can fit curves using linear regression. The most common meth is to include polynomial terms in the linear model. Polynomial terms are independer variables that you raise to a power, such as squared or cubed terms.

To determine the correct polynomial term to include, simply count the number of be in the line. Take the number of bends in your curve and add one for the model order that you need. For example, quadratic terms model one bend while cubic terms moc two. In practice, cubic terms are very rare, and I've never seen quartic terms or high When you use polynomial terms, consider standardizing your continuous independe variables.





Our data has one bend. Let’s fit a linear model with a quadratic term.



The R-squared has increased, but the regression line doesn’t quite fit correctly. The fitted line over- and under-predict the data at different points along the curve. The h R-squared reinforces the point I make in my post about how to interpret R-squared. High R-squared values don’t always represent good models and that you need to che the residual plots!

Let’s try other models.

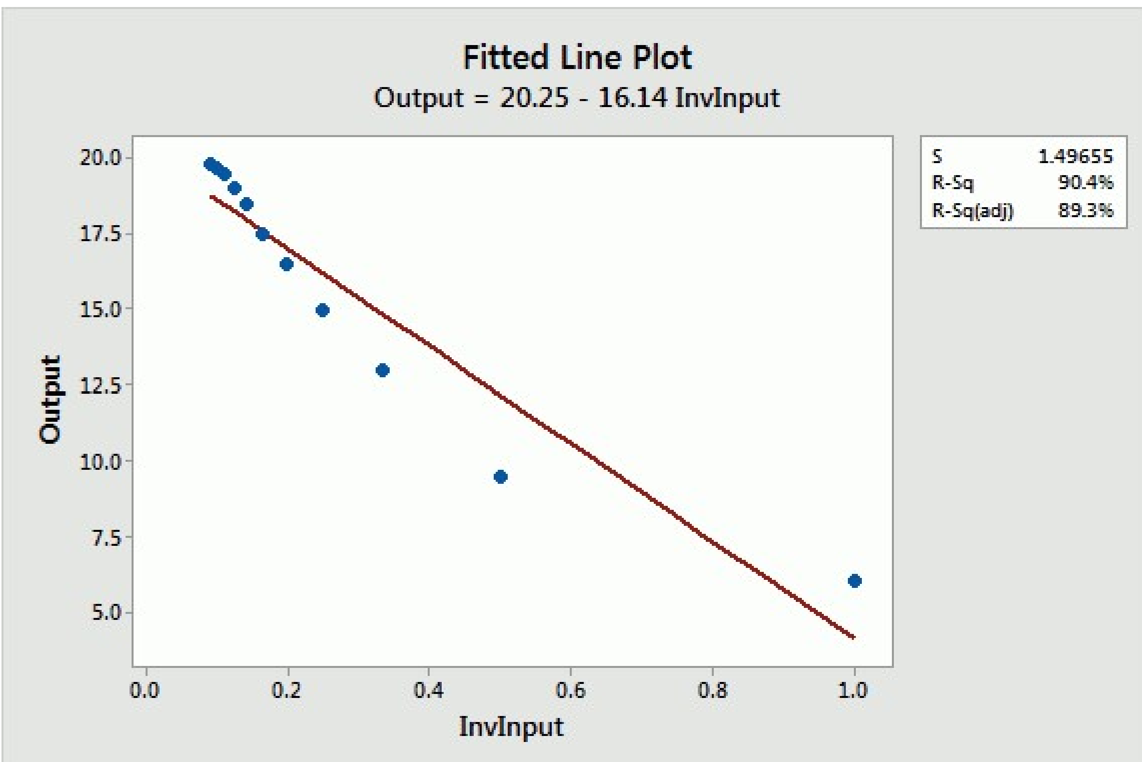
Curve Fitting using Reciprocal Terms in Linear Regression

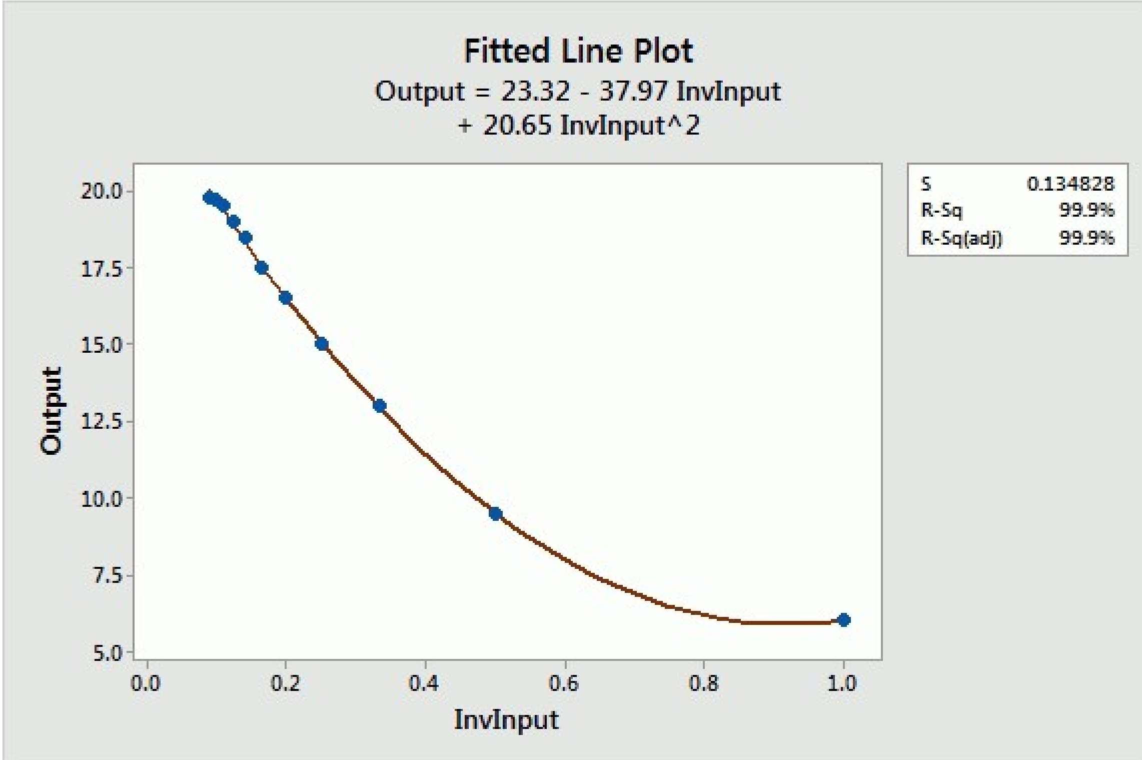
When your dependent variable descends to a floor or ascends to a ceiling (i.e., approaches an asymptote), you can try curve fitting using a reciprocal of an independent variable (1/X). Use a reciprocal term when the effect of an independent variable decreases as its value increases.

$$\frac{\beta * 1}{X}$$

The value of this term decreases as the independent variable (X) increases because i in the denominator. In other words, as X increases, the effect of this term decreases the slope flattens. X cannot equal zero for this type of model because you can't divid by zero.

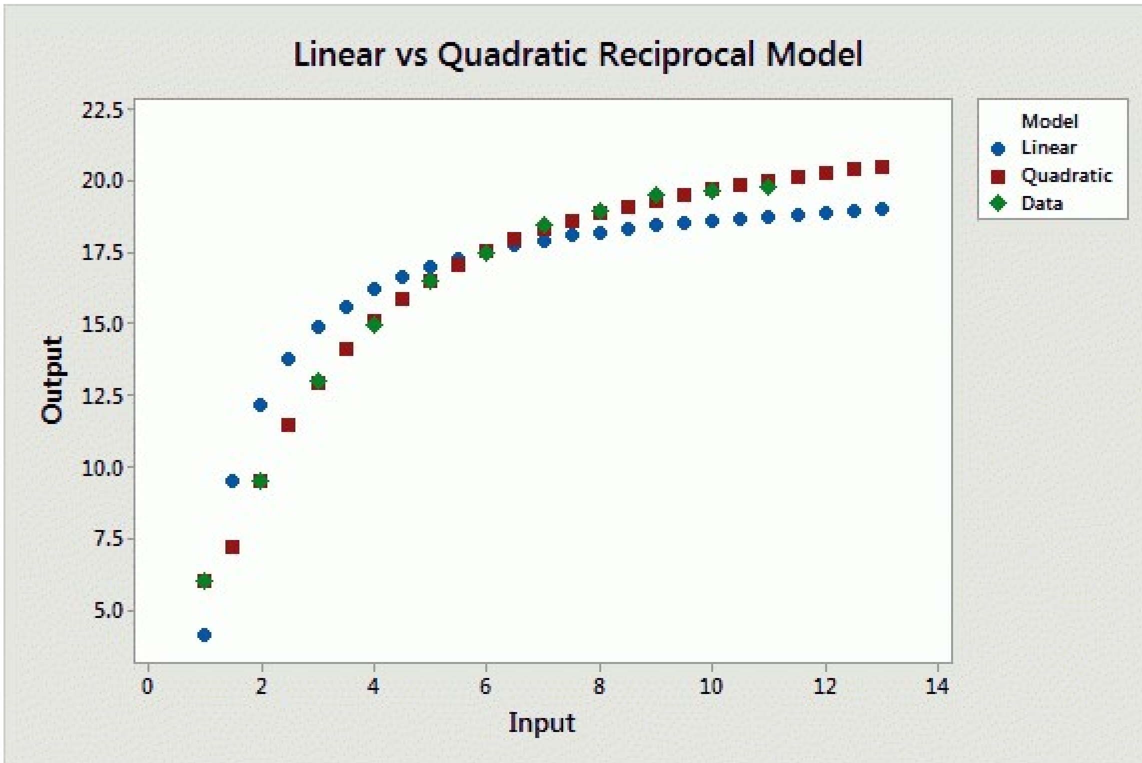
For our data, the increases in Output flatten out as the Input increases. There appea to be an asymptote near 20. Let's try curve fitting with a reciprocal term. In the data I created a column for 1/Input (InvInput). I fit a model with a linear reciprocal term (and another with a quadratic reciprocal term (bottom).





For our example dataset, the quadratic reciprocal model provides a much better fit t the curvature. The plots change the x-axis scale to 1/Input, which makes it difficult to see the natural curve in the data.

To show the natural scale of the data, I created the scatterplot below using the regression equations. Clearly, the green data points are closer to the quadratic line.



On the fitted line plots, the quadratic reciprocal model has a higher R-squared value (good) and a lower S-value (good) than the quadratic model. It also doesn't display biased fitted values. This model provides the best fit to the data so far!

Curve Fitting with Log Functions in Linear Regression

A log transformation allows linear models to fit curves that are otherwise possible only with nonlinear regression.

For instance, you can express the nonlinear function:

$Y=e^{B0}X_1^{B1}X_2^{B2}$

In the linear form:

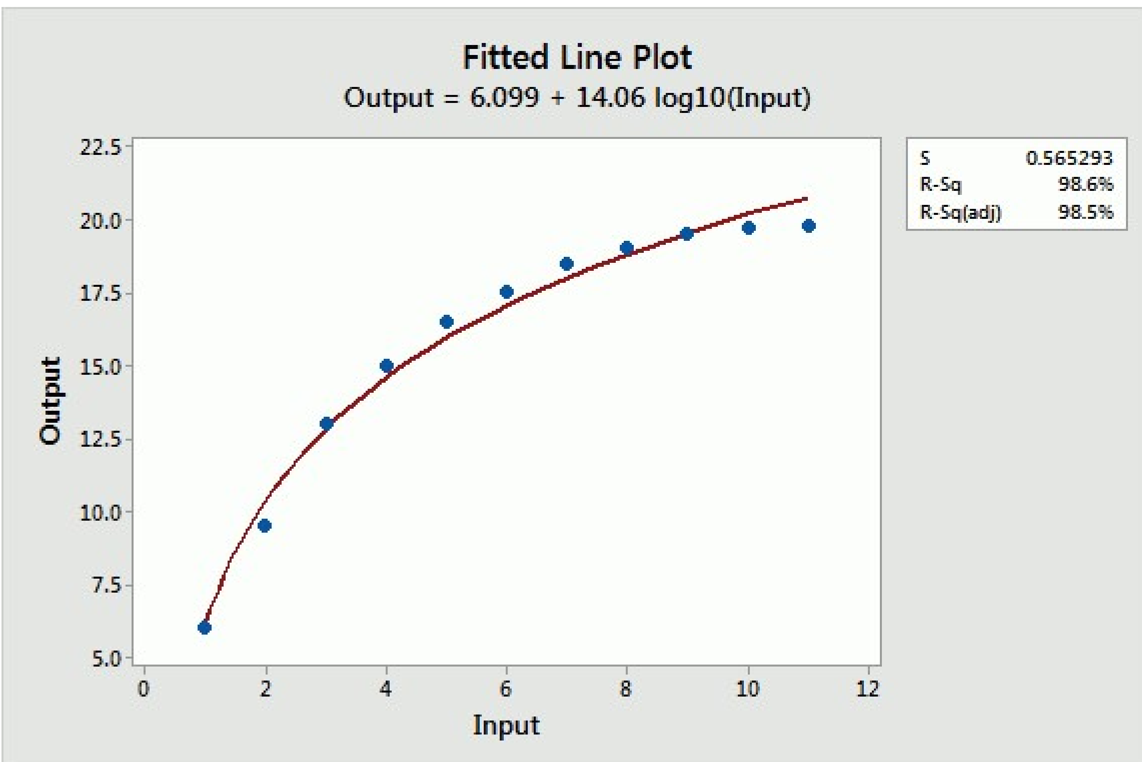
$\text{Ln } Y = B_0 + B_1 \ln X_1 + B_2 \ln X_2$

Your model can take logs on both sides of the equation, which is the double-log forn shown above. Or, you can use a semi-log form which is where you take the log of on one side. If you take logs on the independent variable side of the model, it can be fo or a subset of the variables.

Using log transformations is a powerful method to fit curves. There are too many possibilities to cover them all. Choosing between a double-log and a semi-log model depends on your data and subject area. If you use this approach, you'll need to do s investigation.

Let's apply this to our example curve. A semi-log model can fit curves that flatten as independent variable increases. Let's see how a semi-log model fits our data!

In the fitted line plot below, I transformed the independent variable.



Like the first quadratic model we fit, the semi-log model provides a biased fit to the (points. Additionally, the S and R-squared values are very similar to that model. The model with the quadratic reciprocal term continues to provide the best fit.

So far, we've performed curve fitting using only linear models. Let's switch gears anc
a nonlinear regression model.

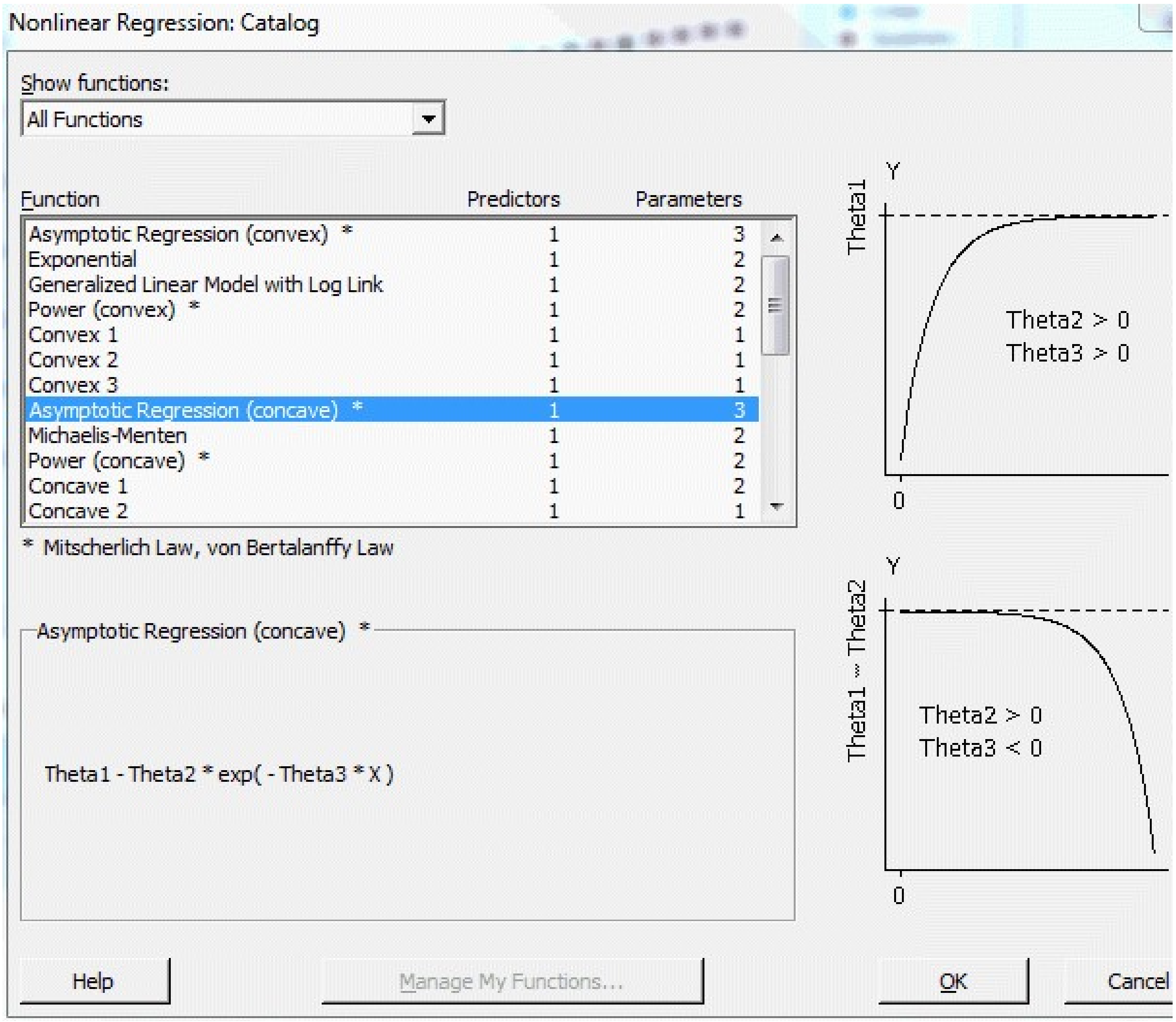
Related post: Using Log-Log Plots to Determine Whether Size Matters

Curve Fitting with Nonlinear Regression

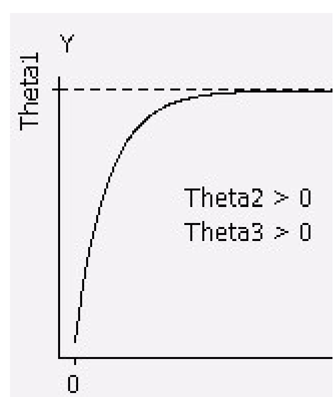
Nonlinear regression is a very powerful alternative to linear regression. It provides more flexibility in fitting curves because you can choose from a broad range of nonlinear functions. In fact, there are so many possible functions that the trick beco finding the function that best fits the particular curve in your data.

Most statistical software packages that perform nonlinear regression have a catalog nonlinear functions. You can use that to help pick the function. Further, because nonlinear regression uses an iterative algorithm to find the best solution, you might need to provide the starting values for all of the parameters in the function.

Our data approaches an asymptote, which helps use choose the nonlinear function from the catalog below.



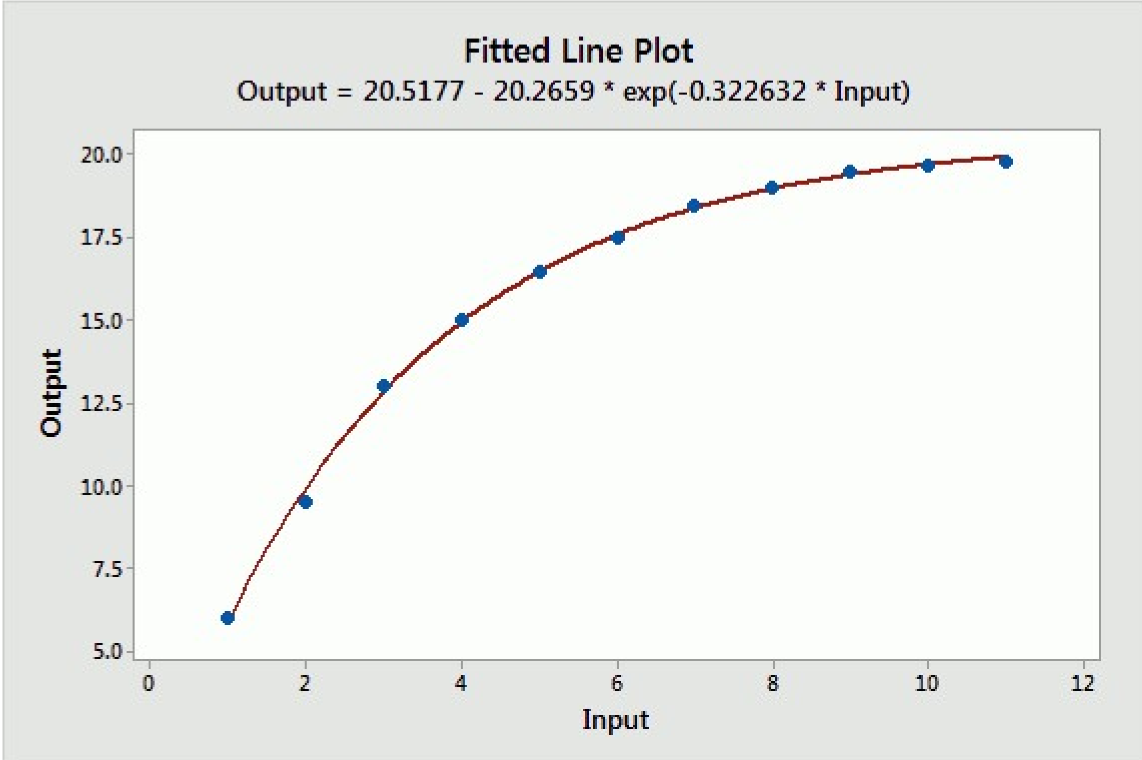
The diagram in the catalog helps us determine the starting values. Theta1 is the asymptote. For our data, that's near 20. Based on the shape of our curve, Theta2 and Theta3 must be both greater than 0.



Consequently, I'll use the following starting values for the parameters:

- Theta1: 20
- Theta2: 1
- Theta3: 1

The fitted line plot below displays the nonlinear regression model.



The nonlinear model provides an excellent, unbiased fit to the data. Let's compare models and determine which one fits our curve the best.

Comparing the Curve-Fitting Effectiveness of the Different Models

R-squared is not valid for nonlinear regression. So, you can't use that statistic to asse the goodness-of-fit for this model. However, the standard error of the regression (S) valid for both linear and nonlinear models and serves as great way to compare fits between these types of models. A small standard error of the regression indicates tl the data points are closer to the fitted values.

Model	R-squared	S	Unbiased
Reciprocal – Quadratic	99.9	0.134828	Yes
Nonlinear	N/A	0.179746	Yes

Quadratic	99.0	0.518387	No
Semi-Log	98.6	0.565293	No
Reciprocal – Linear	90.4	1.49655	No
Linear	84.0	1.93253	No

We have two models at the top that are equally good at producing accurate and unbiased predictions. These two models are the linear model that uses the quadrati reciprocal term and the nonlinear model.

The standard error of the regression for the nonlinear model (0.179746) is almost as low the S for the reciprocal model (0.134828). The difference between them is so sm that you can use either. However, with the linear model, you also obtain p-values fo independent variables (not shown) and R-squared.

For reporting purposes, these extra statistics can be handy. However, if the nonlinea model had provided a much better fit, we'd want to go with it even without those statistics. Learn why you can't obtain P values for the variables in a nonlinear model.

Related posts: The Difference between Linear and Nonlinear Regression Models an How to Choose Between Linear and Nonlinear Regression.

Closing Thoughts

Curve fitting isn't that difficult. There are various methods you can use that provide great flexibility to fit most any type of curve. Further, identifying the best model invo assessing only a few statistics and the residual plots.

Setting up your study and collecting the data is a time intensive process. It's definite worth the effort to find the model that provides the best fit.

Any time you are specifying a model, you need to let subject-area knowledge and th guide you. Additionally, some study areas might have standard practices and functic for modeling the data.

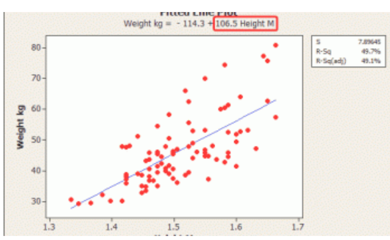
Here's one final caution. You'd like a great fit, but you don't want to overfit your regression model. An overfit model is too complex, it begins to model the random e and it falsely inflates the R-squared. Adjusted R-squared and predicted R-squared ar tools that can help you avoid this problem.

Learn how to choose the correct regression model!

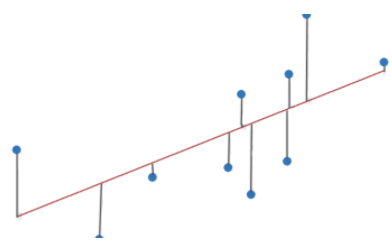
If you're learning regression, check out my Regression Tutorial!

Note: I wrote a different version of this post that appeared elsewhere. I've completely rewritten and updated it for my blog site.

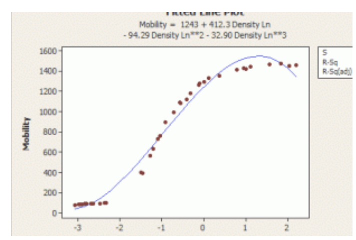
Related



How to Interpret P-values and Coefficients in Regression Analysis
In "Regression"



Choosing the Correct Type of Regression Analysis
In "Regression"



How To Interpret R-squarec Regression Analysis
In "Regression"

Filed Under: Regression Tagged With: analysis example, assumptions, conceptual, interpreting r

Comments

Shaz says
November 15, 2018 at 5:58 am

Hi Jim,

I am getting my head around on understanding one thing. My dependent variable ha lots of zeros. The residuals will potentially be non-normal in this case. But how can t zeros pose a challenge to non-linearity of the relationship? I am confusing the conce of non-linear relationship and non-normal errors in the context of a highly skewed dependent variable with lots of zeros.

Moreover, how can log transformation correct for non-linearity and non-normality here?

★ Loading...

Reply

Jim Frost says
November 15, 2018 at 11:46 am

Hi Shaz,

This is a fairly complicated problem that affects some subject areas more than others. Unfortunately, I don't have any first-hand knowledge of dealing it, which limits how much I can help.

Typically, this type of problem goes beyond using transformation to resolve it.

If you are dealing with count data, you might look into zero inflated models. I discuss those a bit in my post about choosing the correct type of regression analysis. You'll find that in the count data section at the end.

Another method I've heard a bit about is separate your dataset into two dataset One is dataset indicates the presence of whatever you're measuring. The other i the amount. You create separate models for each. Model the presence dataset using logistic regression and the other with ordinary regression. Then, you merg the models That might or might not work for your data.

This issue is something that will probably take a bit of research on your part. Wl write above is really the extent of my knowledge. I'm sure there are also a varie of subject specific variations on this issue as well.

I hope this helps to at least point you in the right direction!

★ Loading...

Reply

cristina says
October 27, 2018 at 7:58 am

how could I fit a nonlinear data set to a linear function?

★ Loading...

Reply

Jim Frost says
October 27, 2018 at 4:32 pm

Hi Cristina,

In the first portion of this post, I show you a variety of ways that you can fit curv using a linear model.

★ Loading...

Reply

Al says

October 7, 2018 at 2:06 pm

Hi jim,

Why does a linear regression model with an x and an x-square term not have high multicollinearity automatically? The correlation between x and x-squared should be high.

★ Loading...

Reply

Jim Frost says

October 7, 2018 at 6:58 pm

Hi Al,

Yes, you're completely correct–and squared terms do cause very high multicollinearity. If you check the VIFs (that measure multicollinearity), you'll fin very high values. Fortunately, there is an easy solution to fix multicollinearity caused by these types of terms. Read my post about multicollinearity for more information!

★ Loading...

Reply

Al says
October 7, 2018 at 2:02 pm

Hi Jim

★ Loading...

Reply

Albert says
August 22, 2018 at 1:56 pm

Hi Jim, Thank you for this thorough explanation!

★ Loading...

Reply

Xie Chang says
August 22, 2018 at 10:09 am

Hello Jim,

I have one question regarding multiple regression. Actually I'm trying to find the ene (E) of an object using the mass (M) and the shape factor (s) multiplied by the velocity as independent variables:

E= β+ β1(M)+ β2 (sV)^2

In this case, I'm using Excel (data analysis: regression option) to find (β, β1 and β2). T best fit (highest R^2) is obtained if the term (sV) is squared. in this case, it is still a

Multiple linear regression or Multiple nonlinear regression because one of the terms squared??

Thanks,

Xie

★ Loading...

Reply

Jim Frost says
August 24, 2018 at 2:39 am

Hi Xie,

It is still linear regression analysis. To learn why, read my post about the differe between linear and nonlinear regression.

Have you tried including the sV term (not squared) as well?

Best of luck with your analysis!

★ Loading...

Reply

Adriano says
July 20, 2018 at 7:43 am

Hi Jim.
Thank for all the strait to the point information.
I have a rather not so simple question, and hoping for a as simple as possible explanation.

I have 10 predictors which affect a specific beer consumption, like: price, trade penetration, advertising, temperature etc.
What is a procedure of fitting a nonlinear regression with more predictors?
Thanks

★ Loading...

Reply

Jim Frost says
July 20, 2018 at 4:07 pm

Hi Adriano,

First off, we need to clarify whether you mean a true nonlinear model or a linea model that uses polynomials to fit curvature. There are huge differences betwe the two types. In fact, I've never heard of a true nonlinear model that has 10 predictors. One seems to be the most common case. So, I'm going to assume th you actually mean a linear model that uses polynomials and/or data transformation. To be sure about this, you should read my post, The Differences between Linear and Nonlinear Models. You'll be able to tell the difference and know what type of model you're using.

As for fitting a model with 10 predictors and potential curvature. Choosing a mo to fit your data is known as model specification. You should read my post about Model Specification: Choosing the Correct Regression Model. This post goes ove the different statistical and non-statistical methods for choosing the best model addition to that information, given that you are particularly interested in model curvature, you should graph the individual relationships between each predicto and the response. This process will help you visually assess curvature and help y include the correct polynomial terms–or possibly use other methods to fit the curve. You should also think about the potential curvature from a theoretical ba These are always important tasks to perform, but more so because you're specifically concerned about curvature.

One final warning. Because you have 10 predictors and possible polynomials, yo
need to worry about overfitting your model. You need a certain number of observations per term in your model or you risk obtaining invalid, misleading results. Read my post about overfitting for more information.

I hope this helps!

★ Loading...

Reply

bob says
June 27, 2018 at 9:22 am

Hi, thanks for your helpful webpage! I'm running some statistic analysis on spss to check for both linear and non-linear effects (about 10 predictor variable and one outcome variable, al are of continiues level) in a multiple linear regression . My goal to check If I can come to a better model for predicting the outcome variable if I chec for posible non-linear effect. I took the folowing steps, Is this a good approach?
-made a linear model with only the significant predictors(function, analyse, regressic linear, "backward", "forward")

-made an extra variable for the ones the literature suggest possible quadratic effect, made new variables by the square of them (so I did a transformation)

-I putted the squared variables in the total model, and checked I they are significant
thanks

★ Loading...

Reply

Jim Frost says

June 27, 2018 at 2:59 pm

Hi Bob,

A quick terminology issue before we get to your question. Linear and nonlinear have very specific meanings in statistics that refer to the form of the model and whether the line is curved. I know that's confusing! That's why I wrote an entire post about that issue–The Difference Between Linear and Nonlinear Regression statistical terms, your model with squared variables is a linear model even thou it will fit a curve!

Your general process sounds correct. Although, I have a few suggestions. For on thing, be sure to assess the residual plots for the model without the squared variables. If there is curvature that you need to fit, you'll often see it in the resid plots. And, those plots are a great way to verify that you're fitting any curvature adequately.

When you include the squared terms, check their p-values to see if they're significant. That can help you determine whether those terms are good additior the model.

Finally, it looks like you're using a stepwise procedure to select your model. Just aware that research shows that stepwise procedures generally only get you clos the best model but not exactly to it. Read my post about Stepwise Regression fo more information. Stepwise chooses the final model based strictly on statistical significance. To specify the correct model, you typically need to use subject-area knowledge and theory to guide you along with the statistical measures. Read my post about Model Specification for more about this!

I hope this helps!

★ Loading...

Reply

Wisley Wan says
May 26, 2018 at 5:39 am

Hi Jim, Please ignore my previous message – I've found it!

★ Loading...

Reply

Jim Frost says
May 28, 2018 at 2:24 am

Hi Wisley,

I'm glad that you find the blog to be helpful. That means a lot to me! I'm also gla
that you were able to try example out yourself!

★ Loading...

Reply

Wisley Wan says
May 26, 2018 at 5:34 am

Hi Jim,

Thank you very much for the blog. It is very clear and helps me understand the issue
better.

I tried the polynomial linear regressions using excel (standardized the IV), but it is w
that the interception is 0 but the other coefficients are both correct. the R-squared

dropped to only 7%. I have checked but couldn't find where went wrong….Could you
please give me some tips?

Thanks!

★ Loading...

Reply

Patrik Silva says
March 29, 2018 at 1:15 pm

It helped for sure!

Thank you Jim, for your prompt answer. I understood very well.

I see all the affection that you are giving us here.

Thank you for sharing your valuable (and I imagine scarce) time with me. I thank you
very much.

★ Loading...

Reply

Patrik Silva says
March 29, 2018 at 11:24 am

Hi Jim, this post is definitely wonderful, because it provide the foundation of regress
analysis…I was always thinking that linear regression is the one where the correlatic
seems to look like linear (line), like you were saying that we maybe think.

Very, very clear!!! However i have some questions:

1) How do we convert back to the original unit of the data, and how can we interpret coefficient, after using the transformation and polynomial terms.

2) By transforming the data for example reciprocal transformation the curve looks inverse, that OK, because that is what we want. But how can we interpret the graph? think we lose all the power to explain the graph since is not in a readable unit. For example if the line has a positive slope we say as the X increase the Y variable tend t increase also. but in the reciprocal everything is inverse, I got lost in this point.

I hope you understand my question and clarify it to me.

Thank you!

Patrik Silva

★ Loading...

Reply

Jim Frost says
March 29, 2018 at 12:10 pm

Hi Patrik,

You're right, the names of the analyses (linear and nonlinear regression) really g the wrong impression about when you should use each one!

On to your questions.

For converting the transformed data back to the original units, you can do the calculations yourself. The precise calculations depend on the nature of the transformations that you've used. However, most statistical software should be able to back convert the values for you. So, I'd check that out first. One thing to note, if you use polynomials to fit curvature, you don't need to back transform anything. All the units use their original scale. For example, suppose your mode y = 2 + 2X + X^2. If your X = 2, then your y = (2 + 4 + 4 = 10). No transformation i

necessary! However, it does make understanding the relationship between the X
and the Y more complex because it changes.

In general, most statistical software can produce main effects plots that incorporate all the transformations. These plots display the relationship betwee an independent variable and the dependent variable while incorporating transformations and polynomials. If the relationship is curved, you'll see it in th graphs. Looking at the graph helps you characterize the nature of the relationsh which brings me to your second question.

For the model that uses the reciprocal, I had to actually create the Linear vs Quadratic Reciprocal Model comparison graph by hand because the software couldn't do that for reciprocal variables. However, once I created the graph, I ca use it to describe the relationship because it's all in natural units at that point.

The way that I'd characterize the quadratic reciprocal relationship is that as inpu increases, the output also increases. Initially the output increases at a very high rate but as input increases, the rate by which output increases slows down as is asymptotically approaches ~20. For example, looking at the quadratic curve in t graph, you can see that increasing X by 1 unit corresponds to different changes If your input is at 1 and you increase it to 2, the output increases by quite a bit– from ~6 to ~10. However, if your input is at 10 and you increase it to 11, the out barely increases at all–stays right around for both settings ~19. Our model incorporates all of that mathematically!

You raise an important point. While these transformations help you fit curves th are present in your data, they can obscure the reality behind the relationships, ' need to transform the numbers back to their natural units and use graphs to understand the relationships. Fortunately, statistical software can automate tha process. One advantage of nonlinear models is that you don't transform the dat but rather you specify the model that fits the data without transformation. However, the relationships, coefficients, etc can be just as hard to understand! / an example of that, just look at the nonlinear model in this post and you'll see tl equation is cryptic! The data aren't transformed but the equation is not easier t(understand. Again, use the graph to better understand the nature of the relationship.

I hope this helps!

★ Loading...

Reply

Ahmed says
March 21, 2018 at 1:11 pm

HI

I have 5 variables with 3 levels and 1 variable for 2 levels. based on that, had designe
18 mixes and i have tested one response for different ages (4 periods).
Now, i have 8 columns: 6 for variables (x1,x2,…x6) and the age, finally the response
I have selected the option of regression- fit regression model and have found the all
anova table and regression model
the problem is the relation between these variables and this response by using mair
effect plot was straight line and this i dont know how can change it to curve
someone told me need to divide the response on the square root of the age

please help me, i appreciate that

★ Loading...

Reply

Jim Frost says
March 21, 2018 at 2:02 pm

Hi Ahmed, you need to fit a model that can handle the curvature, such as by including polynomial terms (e.g., X^2). Based on the analysis names, it sounds li you're using Minitab. If so, include your variables on the main dialog box, then c

Model, and there you can include the higher-order terms (polynomials and interactions). Then, when you create a main effect plot, it should display any curvature that is present.

It sounds like either you're not fitting those polynomial terms or, if you did, may curvature isn't present? How do the residual plots look?

★ Loading...

Reply

Josey says

December 31, 2017 at 7:43 pm

I'm curious. I use non-linear regression to model the progression of prostate cancer The independent variable is PSA (Prostate-specific antigen), a product of healthy and cancerous prostate cells. A half-life regression model works well in predicting wheth the cancer is growing or, conversely, whether the cancer treatment is working. In the former case, cells numbers are doubling. In the latter case, cells numbers are halvin

Is it possible to model both at the same time with the same data? In other words, is possible to estimate how many cells are doubling in number because they are resist to treatment and how many are halving i number because the treatment is effective

★ Loading...

Reply

Leave a Reply

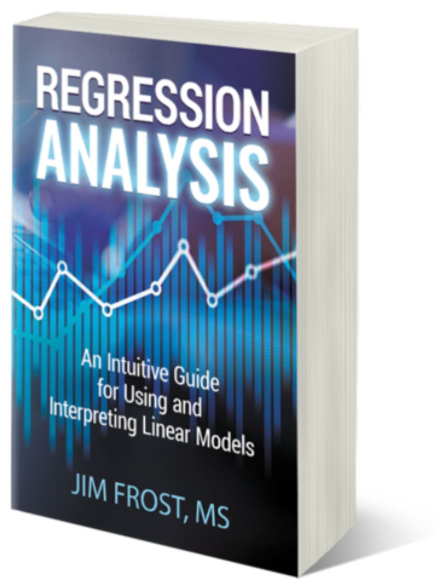
Enter your comment here...

Meet Jim

I'll help you intuitively understand statistics by focusing on concepts and using plain English so you can concentrate on understanding your results.

Read More...

Buy My eBook!



Follow Me

Facebook

RSS Feed

Twitter

Popular **Latest**

- How To Interpret R-squared in Regression Analysis
- How to Interpret P-values and Coefficients in Regression Analysis
- Understanding Interaction Effects in Statistics
- How to Interpret the F-test of Overall Significance in Regression Analysis
- Measures of Central Tendency: Mean, Median, and Mode
- Multicollinearity in Regression Analysis: Problems, Detection, and Solutions
- The Importance of Statistics

Recent Comments

Jim Frost on Heteroscedasticity in Regression Analysis

Luca romen on How To Interpret R-squared in Regression Analysis

phabdallah on Heteroscedasticity in Regression Analysis

Jim Frost on Choosing the Correct Type of Regression Analysis

Syed Abbas on Choosing the Correct Type of Regression Analysis