Mgmt 469

Nonlinear Relationships

Suppose that you have the following variables:

sales: Monthly sales for firm i in year t

adverts: Advertising expense for firm i in year t.

If you type the command in Stata:

regress sales adverts

Stata estimates the following model:

$$Sales = B_0 + B_1 \cdot Adverts$$

By estimating this model you are implicitly assuming that there is a *linear* relationship between advertising and sales:

1) If you plotted the predicted relationship between Adverts and Sales, you would get a straight line.

2) Each additional unit of Adverts is predicted to cause a constant B_1 unit increase in sales.

The assumption of a linear relationship is always a good starting point for analysis. In fact, if your predictors vary over only a narrow range of values, then the linear assumption is very good, as any departure from linearity is probably small.

There are two possible justifications for departing from linearity on the RHS:

1a) Your RHS variable shows considerable variation *and*
1b) You have reason to suspect that the effect of the RHS variable is nonlinear (e.g., you are looking at potential U-shaped cost curves, or perhaps some "dose-response" relationship that shows diminishing returns)

or

2) You are performing robustness checks and want to see if your key conclusions change as you change the model specification

The most common ways to specify a nonlinear RHS variable are with exponents or by taking the *natural log* of the variable.[1] You may also consider departures from linearity on the LHS. Usually, this involves taking the natural log, but it might also involve taking the inverse of the LHS. Most applications involve the log, so this is what we will focus on.

Taking the log of a variable does more than change the shape of a graph. It can profoundly affect how you interpret your results. It is therefore essential to understand the mathematical interpretation of log specifications.

Models with Logs

Empirical researchers usually refer to three possible model specifications:

Linear: The dependent variable and independent variables are untransformed.

Log-log: All continuous variables on the LHS and RHS are logged. (Indicator variables on the right hand side keep their 0/1 values)

Log-linear or **Semi-log**: The dependent variable is logged. All predictor variables remain untransformed.

(Linear-log is possible, but is rarely used in practice)

Each specification has its own unique economic interpretation. You must learn the underlying economics if you are to choose and interpret your models correctly.

[1] Henceforth, when I mention logging a variable, I am referring to the natural log. Econometricians work with natural logs because of the simplicity of the resulting model interpretations, as you will shortly discover.

Why Take Logs?

Some researchers log a variable because it has a long *tail* (it is skewed to the right) whereas its logged value looks like it is normally distributed. These are bad researchers. There is no econometric reason to make a variable appear normal. The only requirement for OLS regression is that *the residuals appear normal*. In your introductory statistics class, you probably examined plots of residuals and log residuals to try to figure which one looks more normal. Alternatively, you may have examined plots of Y on X, log Y on log X, etc., to see which one looks more linear. This is known as the *interocular* approach to empirical research.[2]

Unfortunately, the interocular approach often fails because real life plots of data are rarely as clean as they are in classroom examples. Most of the time, you find yourself staring at clouds of data. It turns out that there is a formal statistical test to determine if taking logs really does improve the fit of your model. We will discuss it soon, but there are more important matters to attend to.

Taking logs affects economic interpretations. Thus, before evaluating the log versus linear choice on statistical grounds associated with the properties of the model's residual, you should evaluate the log versus linear choice on economic grounds associated with the properties of the model itself.

Key point: The choice of a log versus linear specification should be made largely on the basis of the underlying economics. If the economics are ambiguous, then and only then should you perform a statistical evaluation of the residuals.

[2] Otherwise known as eyeballing the data.

You can better grasp how logs affect economic interpretations by considering a simple example, such as the economic relationship between advertising and sales.

- If you specify a linear relationship, then you are implicitly assuming that a *one unit* increase in A causes a B_1 *unit* increase in S.

- It seems just as plausible to suppose that a *one percent* increase in A causes a B_1 *percent* change in S. (You may recall from microeconomics that this implies that the *elasticity* of sales with respect to advertising is B_1.) This is definitely not a linear relationship.

- As we will see, if the latter supposition is appealing, you will need to estimate a log-log model.

- The conclusion: whenever you are choosing between alternative statistical specifications, you are simultaneously choosing between alternative economic models. You had better be happy with your choice of economic model.

A little story telling can help you choose the right economic model. Suppose you are examining the effect of advertising on sales and your sample consists of firms of different sizes and scopes. Small firms (with small product lines and limited geographic scope), find that spending $A on advertising has a small impact on sales. Larger firms may experience proportionately larger increase in sales for the same $A in spending. The following table captures these effects:

Firm	Advertising level	Sales
A	10	500
A	11	525
B	10	2000
B	11	2100
C	10	6000
C	11	6300

If you ran a linear regression of sales on advertising (incorporating firm fixed effects), you would be assuming that a *one unit* increase in advertising generated a constant *B* unit

increase in sales, regardless of the firm. But this pattern clearly does not describe the data very well, as the increase in sales varies dramatically from firm to firm.

Careful inspection does reveal a nice pattern, however. Specifically, it appears that a *10 percent* increase in advertising generates a constant *5 percent* increase in sales. (Thus, the elasticity of sales with respect to advertising equals 0.50.) This makes sense – the effect of advertising should be in proportion to the size of the advertiser. We will soon see that such a pattern can be captured by estimating a log-log specification.[3] Any other specification will fail to accurately uncover what is happening in the data.

You should always start by considering the underlying economics. If you believe that the effect of the right hand side variable on the dependent variable is best thought of in terms of an elasticity, then start with log-log. If you think that the linear relationship is more plausible, then go linear. If you are unsure, then you might want to examine residuals.

A General Discussion about How to Interpret Models with Logged Variables

We can use a bit of math to confirm the intuition that log specifications deal with percentages. Consider the log-log specification. Suppose you have monthly data on sales (S), advertising (A), and price (P). You specify a log-log model:

(1) $$\log S = \log B + \alpha \log A + \gamma \log P$$

You can exponentiate this to obtain another equation:

(2) $$S = B \cdot A^{\alpha} \cdot P^{\gamma}$$

[3] The Stata command for taking logs is **ge logvar=log(var)**. Remember, this is in natural logs.

The coefficients α and β in equations (1) and (2) are identical. Thus, if you insist on estimating the log-log specification of equation (1), then you must believe that the relationship expressed in equation (2) is correct.

Equation (2) states that *the effects of advertising and price on sales are multiplicative,* where the degree of the multiplicative effect is captured by α and γ. For example, regardless of the level of sales, if advertising increases by 10 percent, sales will increase by approximately .10α percent. It bears repeating: if you select a log-log specification, then you are accepting that the effects of RHS variables are multiplicative (and therefore have percentage effects rather than additive effects). The converse applies: if you think the effects of predictors are multiplicative, you should use a log-log specification.

I mentioned previously that the coefficients in a log-log specification can be interpreted as elasticities. Here is the proof. First, differentiate S with respect to A (or P). It is easier to do this with equation (2) (if you don't know how to do this, do not worry about it):

(3a) $$\partial S/\partial A = B \cdot \alpha \cdot A_{it}^{\alpha-1} P_t^{\gamma}$$

(3b) $$= S \cdot (\alpha/A)$$

Equation (3b) implies that $\alpha = (\partial S/\partial A) \cdot (A/S)$, which is the expression for the elasticity of Sales with respect to Advertising.

This bit of calculus proves that if you use linear regression to estimate

regress logsales logadverts logprice

the resulting coefficients are elasticities!

Suppose instead that you have a log-linear regression specification:

(4) $$\log S = B + \alpha A + \gamma P$$

To see how to interpret the results, exponentiate (4) to obtain:

(5) $$S = e^{(B + \alpha A + \gamma P)}$$

where e represents the base of the natural log. Differentiate S with respect to A to yield:

(6) $$\partial S/\partial A = \alpha \cdot e^{(B + \alpha A + \gamma P)} = \alpha \cdot S$$

The effect of a one unit change in A on the value of S is again multiplicative! For example, if $\alpha = .1$, then a one unit increase in A causes S to change by .1S. That is, S increases by 10 percent.

In the case of a linear-log specification, one can use similar methods to show that the effect of a one unit change in the raw (unlogged) RHS variable on the LHS is approximated by $\Delta S = \alpha \cdot (\Delta A/A)$. In other words, a one percent change in A causes an α unit change in S.

The following table summarizes how to interpret different regression models. Note that in all cases involving logs, the results come from the calculus, so they are exactly correct only for small changes.

Model	**Nature of change in X**	**Resulting change in Y**
Linear	One unit change in X	B_x unit change in Y
Log-linear	One unit change in X	$100B_x$ percent change in Y
Linear-log (rarely used)	One percent change in X	B_x unit change in Y
Log-log	One percent change in X	B_x percent change in Y

A Bit of Algebra to the Rescue

When in doubt, you can always rely on algebra to help you interpret the model coefficients. *This is especially helpful for the log-linear specification*. Suppose you have the following log linear model:

$$\text{logsales} = 2 - .06 \cdot \text{price} + .002 \cdot \text{advertising}$$

If price increases by one unit, then logsales will decrease by .06 units.

It is more helpful to express such a change in percentage terms. If the coefficient in a log-linear model is B_x, then the predicted effect of X on Y is as follows:

$$\textit{A one unit change in X causes Y to change by} \quad 100 \cdot (e^{Bx} - 1)\%$$

Thus, when $B_{price} = -.06$, a one unit increase in price is expected to cause sales to fall by $100*(e^{-.06} - 1)\% = 100*(.94 - 1)\% = -6\%$.

If you play around with your calculator a bit, you will find that if B_x is small (say, less than .20 in magnitude), then you can safely state that a one unit change in X will cause a $100B_x$ percent change in Y. The following table shows how to interpret a range of coefficients in log-linear specifications. This table is especially helpful for interpreting coefficients on dummy variables, where a 1 unit change is equivalent to going from a value of 0 to a value of 1. You should confirm some of these values at home, as practice for interpreting the coefficients in your own log-linear models.

Guide to interpreting coefficients in log-linear specifications

Coefficient on X	Effect of one unit change in X (in percentage terms)
.05	5.1% increase
.10	10.5% increase
.20	22% increase
.30	35% increase
.50	65% increase
-.05	4.9% decrease
-.10	9.5% decrease
-.20	18% decrease
-.30	26% decrease
-.50	39% decrease

Making Predictions in a Log Model (optional)

Equation (7) is a general representation of a log-linear model (complete with error term):

(7) $$\log(Y) = \underline{BX} + \varepsilon$$

If you exponentiate both sides of (7) you get:

(8) $$Y = e^{BX} \cdot e^{\varepsilon}$$

If you want to predict the value of Y for any given $\underline{X}$, you must take the expected value of the RHS. Letting $E(\cdot)$ denote the expected value, you obtain:

(9) $$E(Y) = e^{\underline{BX}} \cdot E(e^{\varepsilon})$$

In other words, the predicted value of Y is the exponent of $\underline{BX}$ times the expected value of the exponent of the error term.

Because the expected value of the error is 0, you might think that $E(e^{\varepsilon}) = 1$.[4] If that were the case, then you would conclude that $E(Y) = e^{\underline{BX}}$. To predict Y, you would simply have to

[4] Remember that exp(0)=1

exponentiate BX. Unfortunately, if the errors are skewed, then $E(e^{\varepsilon}) > 1$. You have to account for this when making predictions; you have to compute $E(e^{\varepsilon})$.

Here is how to make predictions from regressions where the LHS variable is expressed in logs and the errors are potentially skewed:

1) Run your regression and recover the predicted values and residuals.

   regress logy x
   predict pred (generates a new variable called pred that equals the predicted value)
   predict resid, residual (generates a new variable called resid that equals the residual)

2) Exponentiate each residual.

   ge expresid=exp(resid)

3) Find the mean of the exponentiated residuals.

   su experesid

The resulting mean is the *adjustment factor*.

   ge adjfactor = xxx (where xxx is the mean that you obtained above.)

4) Multiply your predicted values by the adjustment factor

   ge prediction = pred*adjfactor

Here are a few notes:

1) If the adjustment factor is less than 1.05, you should probably go ahead and ignore the adjustment altogether on the grounds of parsimony.

2) You will need to make the same adjustment in a log-log model.

3) If your model is heteroscedastic, the correct adjustment is substantially more complex.[5] Even so, this simple adjustment is much preferred to ignoring the issue altogether.

[5] We will cover heteroscedasticity at greater length later in the course.

Reporting magnitudes of coefficients in logged heteroscedastic models (optional)

If your log-linear or log-log model is homoscedastic, then you need not worry about making adjustments to the magnitudes of the coefficients. Just use the percentage or elasticity interpretations described above.

In heteroscedastic models, magnitudes must be adjusted to account for any differences in variances of residuals across different values of predictors. This adjustment, known as *smearing*, is fairly straightforward in the case of categorical predictors, as I show below. The adjustment is much more difficult for continuous variables and will not be covered in this class.

1) Suppose you have a dependent variable logY, and a predictor variable X with three categories X=0, X=1 and X=2.

2) You regress logY on X and obtain coefficients B_{X1} and B_{X2}. (Remember, the coefficient on the omitted category is effectively 0.)

3) In a homoscedastic model, you would compute the effect of being in category 2 relative to category 1 as: $\mathbf{exp(B_{x2} - B_{X1}) - 1}$. When comparing category 1 with category 0, you would just report $\mathbf{exp(B_{x1}) - 1}$. Likewise for category 2 versus category 0.

4) If there is heteroscedasticity, and the variance of Y differs in categories 1 and 2, then the **smearing formula** showing the effect of category 2 relative to 1 is as follows:

Effect of category 2 relative to category 1 = $\mathbf{exp[B_{x2} - B_{X1} + .5(Var2 – Var1)] - 1}$

where *Var2 and Var1 are the variances of the residuals for observations in categories 2 and 1* respectively. In a comparison between category 1 and category 0, you would compute

Effect of category 1 relative to category 0 = $\mathbf{exp[B_{x1} + .5(Var1 – Var0)] - 1}$

Again, you would need to make the same adjustment in a log-log model.

This is a lot easier to implement than it first appears. Here is a set of Stata commands and verbal instructions using a categorical variable X and dependent variable logY:

1) **regress logY X**

2) **predict residdavid, residuals** (This computes the regression residual and calls it residdavid)

3) **tabstat residdavid, by(X) stat(var)** (This will show the variance of the residuals for each value of X.)

4) Plug the variances and regression coefficients into the smearing formula

Summary of how to adjust predictions and coefficients in log models:

1) If you are taking the log of the LHS variable, compute the residuals for each observation in the model.

2) Exponentiate the residuals.

3) Compute the mean of the exponentiated residuals. This is the adjustment factor

4) If the adjustment factor exceeds 1.05, then multiply all predicted values of the LHS variable by the factor to obtain the corrected prediction. This prediction is exactly correct in the case of homoscedastic models.

5) You need not adjust coefficients in homoscedastic models.

6) If your model is heteroscedastic, and error variances vary by group, you need to adjust the coefficients to get the correct estimates of being in each group. Use the smearing formula: Effect of X2 relative to X1 = $\exp[B_{x2} - B_{X1} + .5(Var2 - Var1)] - 1$

One More Reason Why Functional Form Matters

In your introductory statistics class, you learned that the choice of functional form can minimize potential heteroscedasticity. You have now learned that the choice of functional form affects the economic interpretation of the results. There is one more reason to worry about functional form, and this is one that you cannot avoid. If you get the functional form wrong, your coefficients may be severely biased. The reason is that when you get the functional form wrong, the errors will be distorted in some systematic way. The computer will try to find predictor variables that somehow relate to the misspecification of functional form. The resulting coefficients on these predictors will be biased. All of this can be very subtle and difficult to identify using the "interocular" approach.

Here is an example taken from real world data[6]. Suppose that you wanted to determine whether foreign cars get better fuel economy than US cars. You have a dummy variable **foreign** and a control for **weight**. You believe that the relationship between **mpg** and **weight** is nonlinear so you add a squared weight term and run the following regression:

regress mpg foreign weight wtsq

You get the following coefficients, all of which are statistically significant:

Dependent variable = **mpg**

weight	.0040
wtsq	1.59e-06
foreign	-2.204
constant	56.53

You conclude, paradoxically, that heavier cars get better fuel economy.

[6] This example is taken from *Getting Started with Stata*.

You present this to an engineering friend who mentions some simple physics to you: The energy required to move 2000 pounds one mile is exactly twice the energy required to move 1000 pounds one mile, all else equal. In other words, there is a linear relationship between **weight** and gallons per mile (**gpm**), which is the inverse of **mpg**. The correct model should therefore be:

regress gpm foreign weight

You run this and obtain the following significant coefficients:

Dependent variable = **gpm**

weight	.0000163
foreign	.00622
constant	-.000735

With the proper model, we now see that heavier cars are indeed less fuel efficient (requiring more fuel per mile). The coefficient on **weight** was biased in the misspecified model.

Model Specification: Letting the Data Help You Decide

In the fuel economy example, there was a sound theoretical reason to invert the dependent variable. Sometimes there is no theory to guide you. When this occurs, you can let the data help you decide which specification best fits the data. One common approach is to plot the (unlogged) dependent variable against the key right hand side variable. If the plot looks linear, then choose a linear model. Next plot the log dependent variable against the RHS variable. If this looks linear, try log-linear. Finally plot log against log; if this looks linear, then run a log-log model. (You might also plot the inverse of the LHS against the linear RHS; this would have revealed a linear relationship between **gpm** and **weight**.)

While this simple approach to choosing log versus linear specifications is widely taught in introductory statistics classes, it has limited real world value. One reason has already been

mentioned – you should rely on economics before you worry about precise fits to the data. But there are two other reasons that have everything to do with fit. First, real world data is not usually so clean. More often than not, you end up staring at clouds! Second, simple two-way plots ignore potentially crucial effects of control variables. If you must let the data dictate the method, it is very helpful to have more quantitative methods to choose among specifications.

Many analysts begin and end by comparing the R^2 of alternative specifications. This is okay, provided that you are comparing linear-linear with linear-log, or log-linear with log-log. In other words, if you are comparing two models with the same LHS variable (and the same observations), it is okay to compare R^2. This provides a useful rule of thumb:

When choosing a model specification, you can compare R^2 among different models, but only if they have <u>exactly</u> the same LHS variables and exactly the same observations.

Comparing models with different LHS variables

Suppose that you have two otherwise identical models, one of which uses a linear LHS and the other uses a logged LHS. *Do not compare R^2!!!* By logging the LHS, you change the potential for a good fit. Consider that a linear regression can get a terrific R^2 if it does a good job of fitting a handful of extreme outliers. If you log the LHS variable, there may be no extreme outliers to fit! The log specification is normally handicapped in comparisons of R^2; it gives you lower R^2 even though it may give better predictions of the unlogged values. In other words, comparing R^2 gives you a pronounced bias in favor of the linear specification, even if the logged model does a better job of prediction. You could exponentiate predictions of the logged model (be sure to make any necessary adjustment as described earlier) and compute the resulting R^2, but there is another approach that is econometrically simpler and superior.

The approach that I will discuss is based on Maximum Likelihood Estimation (MLE). You use MLE to compare the linear and log specifications by performing the *Box-Cox test*.

The Box-Cox Test

The *Box-Cox* test provides a MLE comparison of log versus linear specifications. Some software packages, including Stata, perform the test. I will provide an example in class using a data set containing CEO salaries and job tenure. The simple Stata command is

boxcox salary yrceo

The algorithm takes the dependent variable **salary** and *transforms it* according to the following formula:

$$\text{Box-Cox transformation of salary} = (\text{salary}^{\theta} - 1) / \theta$$

This formula looks formidable, at least until you plug in different values of θ.

If $\theta = 1$, then this is essentially the same as salary. This is a linear specification.

If $\theta = 0$, you have a log specification (the proof is beyond the scope of the class.)

If $\theta = -1$, you have an inverse (1/salary) specification.

The Box-Cox method begins by computing the MLE score when $\theta=1$. (In other words, it runs the OLS regression when the LHS variable is salary and computes the resulting MLE score.) The method then tries other values for θ. Each time, it runs a regression using $(\text{salary}^{\theta} - 1)/\theta$ as the LHS variable and then computes the MLE score. Ultimately, the method reports the value of θ that maximizes the MLE score.

Normally, the best fitting θ is some value other than -1, 0, or 1. If you care only about fitting historical data, you would use this value of θ. But this makes for ugly computations that lack any economic sense. Fortunately, Stata also reports whether it is more appropriate to select θ = -1, 0 or 1. Because you are using this method to help you choose between economically sensible specifications for the LHS, *you should restrict your choice to one of these options*.

Here is a Stata screen for the command **boxcox sales1 price1 promo1** using the yogurtsmall data: Note the best fitting transformation is θ = .59, which is a bit closer to the linear specification. Stata also rejects the null hypotheses that either linear or log fits best (see bottom of Stata screen.) If your goal is to obtain the best possible fit, regardless of the underlying economics, then you should use θ=.59 in the Box-Cox transformation. Of course, this has no economic meaning. Thus, you should probably select θ=1 (linear) as this gives a slightly better fit. (It generates a more positive "log likelihood score": -710 versus -712.

```
                                                Number of obs   =         88
                                                LR chi2(2)      =      82.45
Log likelihood = -707.78448                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
      sales1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .5928201   .1935714     3.06   0.002     .2134272    .9722131
------------------------------------------------------------------------------
Estimates of scale-variant parameters
-------------------------------
             |      Coef.
-------------+-----------------
Notrans      |
      price1 |  -3500.644
      promo1 |   98.25814
       _cons |   487.3508
-------------+-----------------
      /sigma |   27.81587
-------------------------------

-------------------------------------------------------------------------------
        Test |      Restricted     LR statistic      P-Value
         H0: |  log likelihood          chi2       Prob > chi2
-------------+-----------------------------------------------------------------
theta = -1   |     -734.61778          53.67          0.000
theta =  0   |     -712.06208           8.56          0.003
theta =  1   |     -710.13607           4.70          0.030
-------------------------------------------------------------------------------
```

Summarizing Approaches to Selecting a Log or Linear Specification

1) Begin with the underlying economic logic. Do you want to think in linear or percentage terms?

2) If there is no dominant economic logic, let the data speak to you.

3) Select your core group of RHS variables – be sure there are enough so that you are unlikely to improve R^2 very much as you build the model further.

4) Use the Box-Cox test to determine whether to transform the LHS by taking logs.

5) Now that you have settled on the LHS, build the RHS of your model. You may use R^2 to determine whether to use log or linear RHS variables. (Remember that linear-log is rarely used because it is hard to interpret.) You might go back and forth between adding some RHS variables, logging them, taking them out again, etc. Such experimentation is normal. Let robustness be your guide to variable inclusion.

6) Once you have your final model specification, redo the Box-Cox test just to make sure the LHS variable is correctly specified. But do not feel obligated to do Box-Cox every step of the way. This is a waste of time and, if you started with a good core model, totally unnecessary.