Transformations in regression

Everything we’ve done so far assumes a linear relationship between $x$ and $y$. What if that’s not true? Then none of this analysis makes any sense. What are the possibilities? We can determine these from either examination of scatter plots or from our understanding of the underlying process itself.

(1) We might notice a parabolic (quadratic) relationship between $x$ and $y$. This just suggests enriching our model to include both linear and quadratic terms. That is, we should fit a model that includes the two predictors, $x$ and $x^2$. This just falls under the general heading of multiple regression. Note that you should include both $x$ and $x^2$ in your initial model, and usually you would include the $x$ variable in the final model if the $x^2$ variable is needed even if it is not statistically significant, so that a completely general quadratic function is generated.

(2) In some contexts the relationship between $x$ and $y$ is inherently nonlinear. Consider, for example, the science of pharmacokinetics, which is the study of the way drugs spread through the body after being administered to a patient. A standard pharmacokinetic model is the so-called two-compartment model, which says that $C_t$, the concentration of a drug in the bloodstream $t$ minutes after the drug has been administered into a patient’s arm, satisfies

$$C_t = \theta_1 e^{-\theta_2 t} + \theta_3 e^{-\theta_4 t} + \epsilon_t,$$

where $\{\theta_1, \ldots, \theta_4\}$ are parameters that determine the concentration of the drug. The motivation for this model is that the body can be thought of as consisting of two compartments: the vascular system, including the blood, liver, and kidneys, where the drug is distributed throughout the body quickly, and poorly perfused tissues, such as muscle, lean tissue, and fat, where the drug is eliminated more slowly. The only way that this model can fit to observed data is by using nonlinear regression methods. Several statistical packages include such routines (including Minitab), but nonlinear regression estimation is a tricky business. All of the nice properties of linear least squares regression that we take for granted no longer hold for nonlinear regression (e.g., $R^2$ measures can be negative, $t$– and $F$–statistics don’t follow $t$– and $F$–distributions, estimates may be difficult to calculate, the usual confidence and prediction intervals might not be appropriate, etc.). Still, in these circumstances, there is no alternative to the use of nonlinear regression methods. For a discussion of several ways to address...

(3) Certain nonlinear functional relationships are linearizable; that is, they can be converted to a linear form through the use of transformation(s). Two particular models are especially useful:

The log–log model

Consider the functional relationship

\[ y = \alpha x^\beta. \tag{1} \]

This is a multiplicative/multiplicative relationship; it is consistent with proportional changes in \( x \) being associated with proportional changes in \( y \). For example, adding a constant to \( x \) has an undetermined effect on \( y \) in the model:

\[ x \rightarrow x + 2; \quad y = \alpha(x + 2)^\beta = ? \]

On the other hand, multiplying \( x \) by a constant also multiplies \( y \) by a constant:

\[ x \rightarrow 2x; \quad y = \alpha(2x)^\beta = 2^\beta \alpha x^\beta. \]

So, for example, if \( \beta = 1.58496 \), doubling \( x \) is associated with tripling \( y \) (since \( 2^{1.58496} = 3 \)). This functional form is linearizable, since if we take the logarithm of both sides of the equation we obtain

\[ \log y = \log \alpha + \beta \log x; \]

that is, the model is linear after logging both \( x \) and \( y \). This model is particularly appropriate for money data, since money tends to operate multiplicatively rather than additively (think of the rate of return of investments). Examining long right–tailed in the logged scale is often a good idea, as it allows previously indiscernible structure to become more apparent. Finally, the log–log model is often accompanied by heteroscedasticity, which is why taking logs sometimes cures that problem.

The log–log model is also important in the construction and estimation of demand functions. Let \( y \) above represent demand for a product, and \( x \) be the
price. The price elasticity is defined as the proportional change in demand for a proportional change in price; that is,

\[
\frac{dy}{y} = \frac{dy}{dx} = \frac{dx}{x},
\]

where \(dy/dx\) is the derivative of \(y\) with respect to \(x\). Some calculus shows that for the log–log model, the elasticity is a constant \(\beta\), and the log–log model is therefore sometimes called the constant elasticity model. Thus, if it is assumed that elasticities are constant, they can be estimated using the slope coefficient for price in a log–log regression model fit.

The semilog model

Consider the functional relationship

\[
y = \alpha \beta^x.
\]  

(2)

This is a mixed additive/multiplicative relationship; it is consistent with additive changes in \(x\) being associated with proportional changes in \(y\). Adding a constant to \(x\) multiplies \(y\) by a constant:

\[
x \rightarrow x + 2; \quad y = \alpha \beta^{x+2} = \beta^2 \alpha \beta^x.
\]

So, for example, if \(\beta = 2\), adding one unit to \(x\) is associated with multiplying \(y\) by 2. This is sometimes called a semielasticity. This functional form is linearizable, since if we take the logarithm of both sides of the equation we obtain

\[
\log y = \log \alpha + (\log \beta) \ x;
\]

that is, the model is linear after logging \(y\) but not \(x\). This model is particularly appropriate for modeling the growth of objects over time; for example, the total amount of money in an investment as a function of time, or the number of people suffering from a disease as a function of time.

One other context in which a multiplicative model (and hence logging the response variable) is often appropriate is where the response variable is a “survival time”; for example, how long before a mutual fund or hedge fund closes, or the length of time a customer stays with a particular cell phone company (the most common application of such models, of course, is the survival time of a patient
in a clinical trial). A complicating factor in that situation, however, is that the response variable is often censored, in that some observations are still “alive” at the end of the study period (that is, the fund hasn’t failed, or the customer is still with the original phone company). For such observations, all that can be said is that the survival time is at least the observed value. Statistical methods that account for this must be used in this situation, or else estimates of the effects of predictors will be biased. The paper appended to this handout gives an example of such a situation, shows how least squares methods are not appropriate, and describes proportional hazards regression, a method designed to handle censored response values.

When the response variable is in the logged scale, prediction intervals are no longer additive in nature, but are now multiplicative. Assume that logs are in base 10. The rough 95% prediction interval of ±2s for the logged response corresponds to a multiplicative interval in the original scale, since adding 2s to the logged response is equivalent to multiplying the original response by 10\(^2\), and subtracting 2s from the logged response is equivalent to multiplying the original response by 10\(^{-2}\).

Another semilog model
What about fitting a regression where the target variable \(y\) is not logged, but the predictor \(x\) is? This is obviously possible, but the functional relationship it implies between \(y\) and \(x\) is a little strange:

\[
10^y = \alpha x^\beta
\]

(I’ve used 10 as the base here, assuming that the logs being taken are to that base). Logging both sides gives the relationship

\[
y = \log \alpha + \beta \log x.
\]

Interpretation of the slope \(\beta\) comes from the usual interpretation, except that adding one to \(\log x\) corresponds to multiplying \(x\) by 10. That is, the model implies that multiplying \(x\) by 10 is associated with an expected increase of \(\beta\) in \(y\) (in a multiple regression, holding all else fixed); that is, a multiplicative/additive relationship. When might such a relationship make sense? It seems most appropriate in the situation where the target variable is a “pure” number of some sort, such as a return or a score of some sort, and the predictor is long right–tailed.
Note that for all of these models, you should always interpret all of your regression results in terms of the original variables, not in terms of the logged scale. Thus, you should not interpret coefficients in terms of changes in a logged predictor or response, but rather in terms of the original variables (e.g., as elasticities or semielasticities). So, for the log-log model a slope coefficient \( \hat{\beta}_j \) should always be interpreted as “a 1% change in \( x_j \) is associated with a \( \hat{\beta}_j \)% change in \( y \) holding all else in the model fixed” (or, if you prefer, “multiplying \( x_j \) by 10 is associated with multiplying \( y \) by \( 10^{\hat{\beta}_j} \) holding all else in the model fixed” [assuming that logs base 10 were used]); for the first semilog model a slope coefficient \( \hat{\beta}_j \) should always be interpreted as “a 1 unit change in \( x_j \) is associated with a \( 10^{\hat{\beta}_j} \) multiplicative change in \( y \) holding all else in the model fixed” (assuming logs base 10 were used, and using the actual appropriate units for \( x_j \)); and for the second semilog model a slope coefficient \( \hat{\beta}_j \) should always be interpreted as “multiplying \( x_j \) by 10 is associated with a \( \hat{\beta}_j \) unit change in \( y \) holding all else in the model fixed” (assuming logs base 10 were used, and using the actual appropriate units for \( y \)). Similarly, if the response is in the logged scale, a rough prediction interval should not be reported as “±2s logged units,” but rather “an interval corresponding to predicting the response to within a multiplicative factor of \( 10^{2s} \), 95% of the time” (assuming logs base 10 were used).

An issue that often comes up when dealing with variables that ultimately need to be logged, even though it is not directly related to taking logs, is when variables should be analyzed in a standardized (rescaled) form. If, for example, you looked at the relationship between the number of people who are born and the number of people who die in a given state in a particular year for the 50 states, you are likely to see two dominant patterns: the variables are right-tailed, and there is a direct relationship between the two variables. The former pattern suggests logging the two variables, which will probably make the relationship between them more consistent with the assumptions of least squares linear regression. Unfortunately, this is a very uninteresting relationship, and is not really worth thinking about, because (almost) all of what you are seeing is just a population effect: states like California, Texas, and New York have a larger number of both births and deaths because they are heavily-populated states, and states like Alaska and Wyoming have a smaller number of both births and deaths because they are lightly-populated states. This population effect is not of any importance; what is interesting is whether birth rates and death rates are related, as that could have sociological or policy implications. Whenever you see a variable that is directly driven by population, such as total GDP, total consumption, total production, numbers of people born, died, married, divorced, etc., these should be

© 2015, Jeffrey S. Simonoff
treated in a per person scale (GDP per capita, consumption per capita, production per capita, birth rate, death rate, marriage rate, divorce rate, etc.). If you then need to log the rescaled variable that’s fine, but you will avoid misinterpreting an uninteresting population effect as something much more meaningful. Note also that using population as a predictor of a per capita measure can be quite meaningful: you could discover, for example, that more populous states have higher per capita incomes or lower murder rates, which would be interesting from a sociological or economic point of view.

One thing you might wonder about is “mixing-and-matching” predictors; that is, can you build a model that includes both logged and unlogged predictors? Simply stated — sure. The interpretations given above are still valid as appropriate, as long as you remember to say “holding everything else fixed.” So, if your response variable is logged, coefficients for logged predictors are (partial) elasticities, while coefficients for unlogged predictors are (partial) semielasticities. If your response variable is unlogged, the coefficients for unlogged predictors have the usual additive/additive interpretation, while those for logged predictors have the multiplicative/additive relationship noted earlier. Note, however, that including both logged and unlogged versions of the same variable in the same model (that is, both \( x \) and \( \log x \)) corresponds to fitting a strange nonlinear model (not one of the models mentioned here), and it is difficult to imagine any situation where that would be appropriate.

Note that a situation where a log-log or semilog relationship is appropriate from a contextual point of view, but the relevant predictor and/or response variable has many zero or negative values, is one where directly fitting models (1) – (3) using nonlinear regression could be appropriate.

**Minitab commands**

To fit a quadratic model, just create the squared variable and fit a multiple regression model. Click on **Calc \( \rightarrow \) Calculator.** Enter the new variable name under **Store result in variable** and \( X^2 \) under **Expression** (if \( X \) is the name of the variable).

To take logs (base 10), click on **Calc \( \rightarrow \) Calculator.** Enter the new variable name under **Store result in variable** and \( \text{LOGT}(X) \) under **Expression** (you can also get it by double-clicking on **Log base 10** under **Functions:**). Natural logs are taken using the **LN(X)** function.