Transforming the response and/or predictor variables has the potential to remedy a few of the assumption problems. We can transform either the response variable Y or the predictor variable X, or of both.

Keep in mind that data transformation is a 'trial and error' sort of approach. In building a model, you need to attempt a transformation to determine if the transformation has eliminated the problems with the model. If it doesn't, try another transformation. Continue until you have an appropriate model.

Always remember that there is often more than one viable model. The model you choose and the model someone else chooses may be different, but they may both be equally appropriate. What's important is that your model

- is not overly complicated
- meets the assumptions of linear regression, and
- allows you to answer the research question at hand

Transforming Only the Predictor(s)

– The first transformation technique to consider is transforming only the predictors, the X values. You should consider this type of transformation when the error terms are reasonably close to being normal and are homoscedastic. A transformation on Y is not advised in this situation because transforming Y can materially change the shape of the distribution of the error terms and can lead to the variances shifting around too much. All of this means that you should consider transforming the predictors when linearity is the only issue.

– Figure 1 contains some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful in linearizing the regression relationship without affecting the distributions of Y.

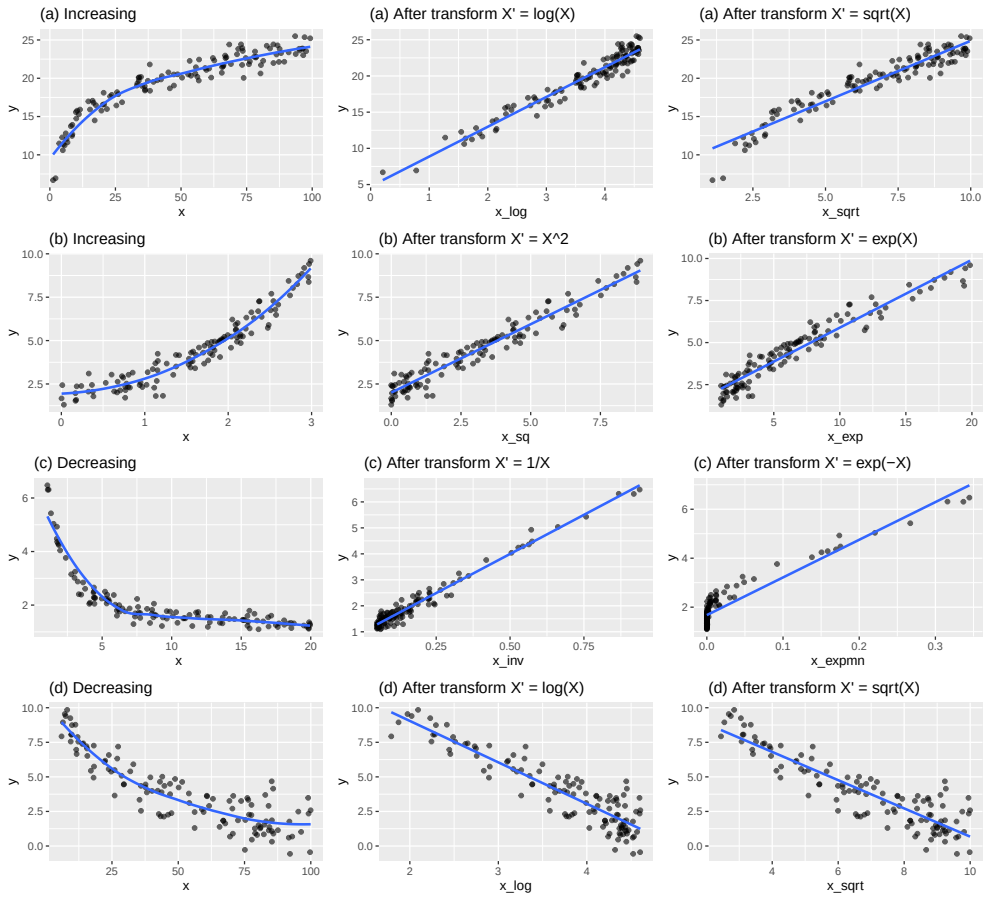

Figure 1: Prototypical nonlinear regression patterns and suggested transformations

Example 1

Data from an experiment on the effect of number of days of training received on performance in a battery of simulated sales situations are presented in the table below, for the 10 participants in the study. The data are given in the `Performance.csv` file.

a. Construct a scattplot on the SLR fit on the raw data. Determine what kind of transformation is appropriate, if any, and perform that transformation. Then, recheck each of the regression assumptions.

b. Write down the transformed regression line. And find the predicted performance score for someone that has a two-day training.

c. Construct and interpret a 95% confidence interval for the average performance score for someone that has a two-day training.

Example 2

Consider a dataset from a memory retention experiment in which 13 subjects were asked to memorize a list of disconnected items. The subjects were then asked to recall the items at various times up to a week later. The proportion of items (y) correctly recalled at various times (x, in minutes) since the list was memorized were recorded. The data are given in the `WordRecall.csv` file.

a. Determine if a transformation on the predictor only is appropriate. Fit an appropriate transformation. Recheck the regression assumptions and determine the quality of the fit of the transformation.

b. What proportion of words can we expect a person to remember after 1000 minutes? Also, construct and interpret a 95% confidence interval.

c. How much do we expect the proportion of words remembered to decrease if the minutes increases by a factor of 10?

Transforming Only the response

– The second transformation technique will involve transforming only the response (Y) values. You should consider this type of transformation if you have heteroscedasticity (unequal variances) and non-normality of the error terms. These two problems tend to happen together. Transforming the Y should be considered to help remedy these problems. Additionally, a transformation on Y may help straighten out a curvilinear relationship.

– Frequently, the non-normality and unequal variances of the errors take the form of increasing skewness and increasing variability of the distributions of the error terms. An example could be household expenditure for vacations (Y) and household income (X). There tends to be more variation and right-skewness for high-income households than for low-income households, who tend to spend much less on vacations in general.

Box Cox Transformation

– It is often difficult to determine which transformation of Y is the most appropriate when observing the diagnostic plots. There is a special class of transformations called the Box-Cox transformations that can help determine the best way to approach transforming data. The Box-Cox transformations automatically identify a transformation from the family of power transformations on Y. A power transformation is a family of functions applied to create a monotonic transformation using power functions. The family of power transformations is of the form

$$Y' = Y^{\lambda}$$

The general form of the Box-Cox transformation is

$$g(y_i) = \frac{y_i^{\lambda}}{\lambda}$$

λ is a parameter that is determined from the data itself.

– There are some observations that need to be made concerning $g(y_i)$ for specific values of λ:

- If $\lambda \approx 1$, then no transformation is needed. The original data should be modelled.
- If $\lambda \approx 0$, then in the limit as λ converges to 0, the Box-Cox transformation is the **log** of the original response variable.
- If $\lambda \approx 2$, then the Box-Cox transformation is the **square** of the original response variable.
- If $\lambda \approx -1$, then the Box-Cox transformation is the **reciprocal** of the original response variable
- If $\lambda \approx 1/2$, then the Box-Cox transformation is the **square root** of the original response variable
- If $\lambda \approx -1/2$, then the Box-Cox transformation is the **reciprocal of the square root** of the original response variable.

– The normal error regression model with the Box-Cox transformation becomes

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i$$

– R gives the value of λ automatically; however, the process for finding it is pretty fascinating. It involves the maximum likelihood method to get the value $\hat{\lambda}$, which is the MLE of λ to use in the Box-Cox transformation.

– It is often reasonable to use a nearby λ value for which the power transformation is easy to understand. For example, use $\lambda = 0$ instead of say $\lambda = .13$, or use $\lambda = -.5$ instead of $\lambda = -.79$ to help facilitate understanding without sacrificing much in terms of the effectiveness of the transformation.

Example 3

Data on age (in years) and plasma level of a polyamine for a portion of 25 healthy children was carried out. The data are given in the `PlasmaLevel.csv` file.

a. Determine what transformation would be most appropriate in this case, if any, and perform that transformation. Then, recheck each of the regression assumptions.

b. What's the expected plasma level of a child with an age of 2.34 years? Construct and interpret a 95% confidence intervals.

c. What's the expected change in plasma level when age increases by one year?

Example 4

The `Mammal.csv` dataset considers the typical birth weight (in kg) and length of gestation (in days) for various mammals.

a. Determine if a transformation on the response is appropriate. Fit an appropriate transformation.

b. What's the expected gestation length of a new 50 kg mammal? Construct and interpret a 95% confidence intervals.

c. What's the expected change in gestation length for each kg increase in birth weight?

Transforming both the response and the predictor(s)

– As you can imagine, this technique is reserved for when the regression function is not linear and the error terms are not normal and have unequal variances. In general (NOT ALWAYS!),

- transforming the y values can correct problems with the error terms (and may help a bit with non-linearity)
- transforming the x values can correct problems with the linearity

– Typically, you want to investigate the linearity first. It's hard, if not impossible, to really investigate homoscedasticity if linearity is not met. However, it typically doesn't matter which you transform first, unless you decide to do a Box-Cox transformation since the λ value depends on the current model at hand. For example, fitting a Box-Cox to $y = \beta_0 + \beta_1 x$ will typically yield a different λ to $y = \beta_0 + \beta_1 \log(x)$.

Example 5

Many different groups, such as the lumber industry, ecologists, and foresters, benefit from being able to predict the volume of a tree using only its diameter. The `ShortLeaf.csv` dataset contains the diameter (in inches) and volume (in cubic feet) of 70 shortleaf pines.

a. Fit a simple linear regression model relating volume to diameter. Construct the diagnostic plots to check the regression assumptions and comment on their validity.

b. Refit the model using a log transformation on both the response and the predictor. Comment on whether the regression assumptions appear reasonable for this transformed model.

c. Using the log-log model, test whether there is a significant association between diameter and volume at the $\alpha = 0.05$ significance level.

d. Based on the log-log model, estimate the average volume of all shortleaf pine trees with a 10-inch diameter.

e. Using the log-log model, determine the expected change in volume for a two-fold increase in diameter.

Example 6

The `Advertising.csv` dataset contains information on product sales (in thousands of units) and advertising budgets for TV (in thousands of dollars).

a. Fit a simple linear regression model relating sales to TV advertising. Write down the fitted regression equation and report the R^2 value.

b. Calculate the predicted change in sales when TV advertising increases from \$5,000 to \$10,000.

c. Find the predicted sales when the TV budget is \$100,000. Construct a 95% prediction interval for this estimate.

d. Create the standard diagnostic plots and comment on whether the model assumptions appear reasonable.

e. Apply a log transformation to both variables and repeat parts a. through d. using the transformed model.

f. Compare the two models (original and log-log). Which model would you choose as the final model, and why?