Class notes 4

Functional form

Non-linearity

Consider the general linear multiple regression model given by

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + \varepsilon$$

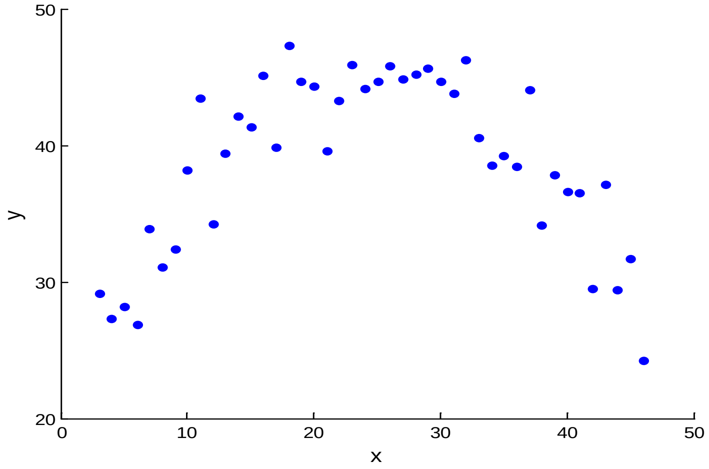
This model implies that the relationships between Y and the independent variables are linear. However, it is often the case that the relationships are non-linear. Under such conditions, it is wrong to force a linear model on the data. In many cases, we can handle non-linear relationships by transforming the variables in the model. There are many forms of non-linearity. In this set of notes, we will discuss how to handle the most obvious ones.

Sometimes there will be theoretical arguments for a specific functional form, which in turn dictates how and if the variables in the model should be transformed. In cases where we cannot rely on theory, one may use empirical methods to detect the likely form of the relationship between the dependent variable (Y) and the independent variables (the Xs). An obvious empirical approach is by visual inspecting a scatter plot between the dependent variable and the independent variable.

Below, we study a few examples

Polynomial models

Let $\{y_i, x_i\}_{i=1}^{n}$ denote a set of n pairs of observations on Y and X. Consider the following plot indicating the relationship between the two variables



Clearly, the relationship between the two variables appears to be quadratic. It would therefore be incorrect to specify a linear model of the form

$$Y = B_0 + B_1X + \varepsilon$$

A more appropriate specification would be

$$Y = B_0 + B_1X + B_2X^2 + \varepsilon$$

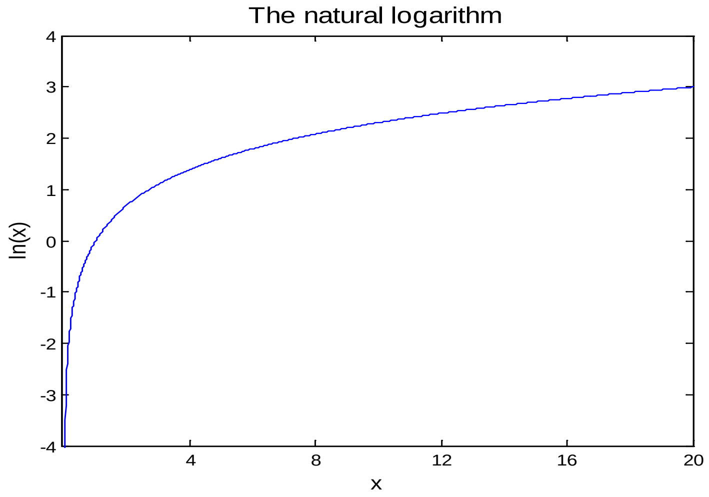
The nature of the relationship showing in the plot additionally suggests that $B_1 > 0$ and $B_2 < 0$. Note that the interpretation of B_1 and B_2 is not as straightforward as in the linear case. The quadratic model belongs to a class of specifications known as polynomial models.

Within this class, relationships involving terms of higher order terms are also possible. For instance,

$$Y = B_0 + B_1X + B_2X^2 + B_3X^3 + \varepsilon$$

Logarithmic models

A popular form of transformation is the logarithmic transformation. The natural logarithm of the (mathematical) variable X is a function written as $\log(X)$ where X is a number larger than zero. The illustration below shows the value of $\log(X)$ for values of X up to 20.



Interpretation of the model parameters

The natural logarithm has a number of desirable properties that makes it popular in regression analysis. One of those properties is that it allows for an easy interpretation of the slope parameters in the model. Consider the (semi-log) model

$$\log(Y) = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k + \varepsilon$$

In this model, $100 \cdot B_1\%$ represents the (approximate) percentage change in Y for a one unit increase in X_1, holding $X_2, \ldots, X_k$ fixed (or constant). Similar interpretation applies to $B_2, \ldots, B_k$.

Next, consider the (double-log) model

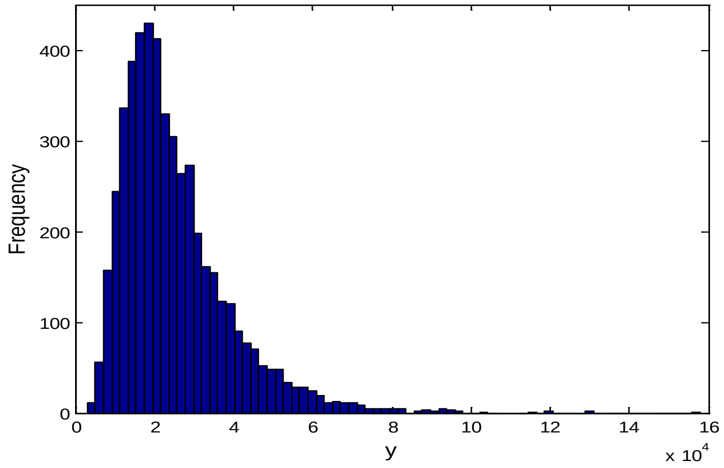
$$\log(Y) = B_0 + B_1 \log(X_1) + B_2X_2 + \cdots + B_kX_k + \varepsilon$$

The parameter $B_1\%$ represents the (approximate) percentage change in Y for a one percent increase in X_1, holding $X_2, \ldots, X_k$ fixed (or constant). Note that $X_2, \ldots, X_k$ are not subject to any transformation. Thus, the interpretation of $B_2, \ldots, B_k$ remain the same as for the semi-log model.

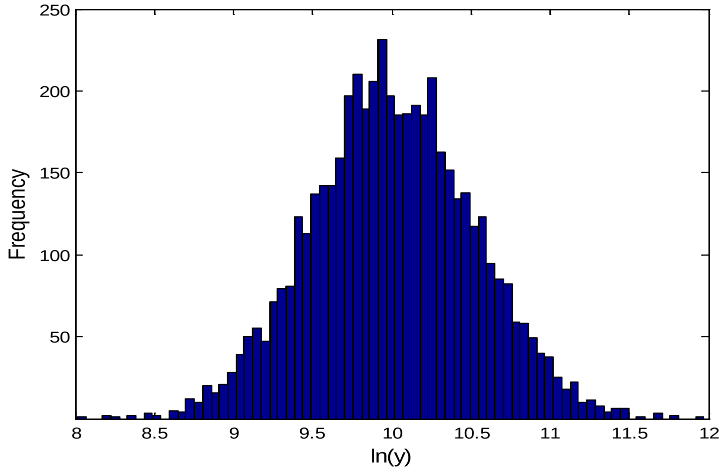
When should we apply the logarithmic transformation?

Since the natural logarithm is not defined for values less than or equal to zero (at least not within the context of these notes), it is not appropriate to apply this form of transformation unless the variable in question is strictly positive for all data points. It may not be obvious when to use logarithmic transformation. Typically it is applied when the considered variable is a *price* variable such as *firm sales*, *firm value*, *consumption*, *savings*, *income* (for instance, *wages*, *GDP*) etc. It is also quite common to transform so-called *count* variables. One problem is that such variables may take the value zero. There are specialized regression techniques aimed at modeling *count* variables.

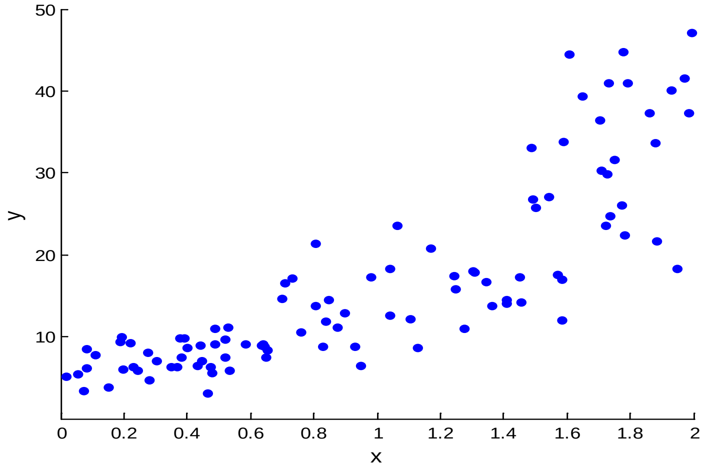
Strictly positive variables are often skewed, and may be severely so. Applying logarithmic transformation will often lead to a more symmetric distribution. As an example, let $\{y_i, x_i\}_{i=1}^{n}$ denote a set of n pairs of observations on Y and X. Now, consider the empirical distribution (histogram) of Y illustrated below



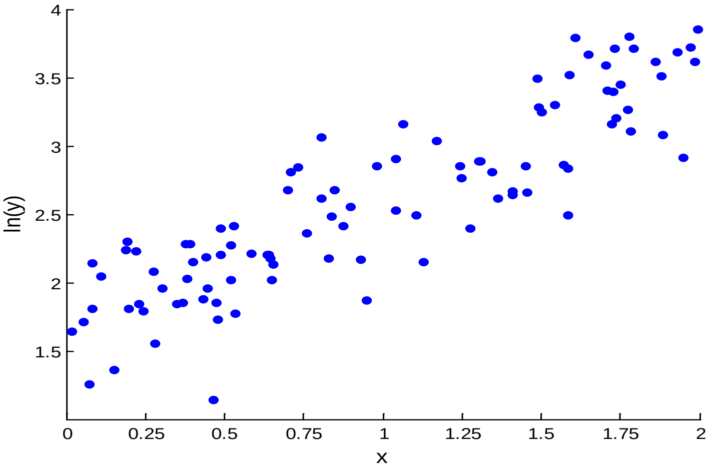
By taking the natural logarithm of the data the points, one obtains an empirical distribution that is more symmetric as illustrated below



The observed skewness in Y seen in the first histogram may be due to the functional relationship between Y and X. Based on the data, we are able to investigate the relationship between two variables by plotting the data points



The scatter plot reveal two things. First, the relationship between Y and X appears to be non-linear. Second, the variance of Y seems to increase with the value X (as X increases, more variation along the y-axis is observed). This phenomena is known as heteroscedasticity. Due to the non-linearity, it is wrong to fit a linear model to the data. To see this more clearly, consider a plot of the same data after applying the log-transformation to Y:



Clearly, the transformation improves the linearity and removes the heteroscedasticity that was seen in the previous plot. This is an example of how non-linearity can be detected and how to correct it.