Log Transformation in Simple Linear Regression (1/3)

a_statistician

February 24, 2017

The simple linear regression model can be write as:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where i indexes the subject and $i = 1, 2, ..., n$ and n is sample size. Y is response (dependent) variable, X is covariate (independent variable), α is intercept, β is slope, and ϵ is error term which is random and following a certain distribution.

The validation of the statistical inference on α and β and the function of them depends on the assumption that ϵ follows a normal distribution. If $\log(Y)$ and X have a linear relation, it is possible to tempt the analysts to fit the simple regression between $\log(Y)$ and X. But this approach may destroy the property of normal distribution of error term.

Here is a dataset in simple ASCII format and you can download it. After downloading the data, you can use the linked SAS program to get the mentioned graphs by yourself if you have SAS software. The scatter plot of variable Y vs variable X indicates that $\log(Y)$ and X have a linear relation and the scatter plot of variable $\log(Y)$ vs variable X verifies it. After fitting the simple regression model between $\log(Y)$ and X, the residuals, which are the estimates of the error terms, are not normally distributed as displayed by the histogram. In fact, Y and $\exp(X)$ also have a linear relation, After regressing Y on $\exp(X)$, the residuals seem very close to normal distribution. Therefore, for this dataset, if the simple linear model is applied to $\log(Y)$ and X, the confidence intervals, statistical tests for α and/or β and their functions are problematic, especially when the sample size is not large. Contrariwise, if the simple linear model is applied to Y and $\exp(X)$, the validation of the statistical tests for α and/or β and their functions is strengthened a lot, because the residuals are very close to a normal distribution.

Because the distribution of error terms is the base of the simple linear regression, we need to make the decision on transformation of response variable based on the empirical distribution of the residuals, instead of on the relationship between Y and X.