Log Transformation in Simple Linear Regression (2/3)

a_statistician

March 3, 2017

The simple linear regression 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 the sample size. Furthermore, Y is the response (dependent) variable; X is the covariate (independent variable); α is the intercept; β is the slope; and ϵ is the error term which is random and follows a specified 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.

Sometimes, the statisticians make the log transfomation on response variable Y, because they find that $\log(Y)$ follows a normal distribution. Does this kind of decisions make sense? Let see following example.

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 histograms indicates that Y is positive skewed distribued and $\log(Y)$ is very close to a normal distribution. In addition, the hrefhttps://drive.google.com/open?id=0B65GsGBz-UakeGkyZjZ5X2dOZ0Uscatter plots dsiplay the linear relation between Y and X, so after transformation of Y, the transformaiton of X is needed to keep the linear relation between response variable and covariate. Therefore the smple linear regression using $\log(Y)$ as response variable and $\log(X)$ as covariate is fit. Let check the estimates of error tems — residuals. The histogram of the residuals tell us that residuals are negative skewed distributed andf are far from a normal distribution. Now we try to fit another simple linear regression, Y on X without any transformation. As regular approach, we check the resulduals again. From the histogram of the residuals, it is hard to tell the difference between this histogram and fitted noraml distribution.

Because the distribution of error terms is the base of the simple linear regression, we should use the simple linear regression Y on X, instead of $\log(Y)$ on X, although $\log(Y)$ is very close to a noraml distribution.

When talking about the assumption of normal distribution for simple linear regression, we generally say that response variable $Y_i, i = 1, 2, ..., n$ is normally distributed. In fact,

$Y_i, i = 1, 2, ...n$ do not follow a SINGLE normal distribution. And $Y_i, i = 1, 2, ...n$ follow m normal distributionS, where m is number of unique values among $X_i, i = 1, 2, ..., n$, given that $\beta \neq 0$, because the Y_i and Y_i' have different means $\alpha + \beta X_i$ and $\alpha + \beta X_i'$ if $X_i \neq X_i'$. So the distribution of $Y_i, i = 1, 2, ..., n$ is the mixture of the normal distributions and this distribution is uncheckable.

Combining the another example, it is clear that log transformation of response variable Y should be determined by the distribution of residuals, instead of the relationship between Y and X or the marginal distribution of Y and/or $\log(Y)$. This conclusion is correct for other kind of transformation of response variable and other models covered by general linear model thoery, such as ANOVA, multiple linear regression model.