



StatNews #83

Interpreting Coefficients in Regression with Log-Transformed Variables¹

June 2012

Log transformations are one of the most commonly used transformations, but interpreting results of an analysis with log transformed data may be challenging. This newsletter focuses on how to transform back estimated parameters of interest and how to interpret the coefficients in regression obtained from a regression with log transformed variables. A log transformation is often useful for data which exhibit right skewness (positively skewed), and for data where the variability of residuals increases for larger values of the dependent variable. When a variable is log transformed, note that simply taking the anti-log of your parameters will not properly back transform into the original metric used.

To properly back transform into the original scale we need to understand some details about the log-normal distribution. In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. More specifically, if a variable Y follows a log-normal distribution, then we have that $\ln(Y)$ follows a normal distribution with a mean $= \mu$ and a variance $= \sigma^2$. Given that, here are the important properties of the log-normal distribution in terms of the original variable Y :

- Mean of variable $Y = e^{\mu + \frac{1}{2}\sigma^2}$
- Median, or geometric mean of variable $Y = e^{\mu}$
- Variance of variable $Y = (e^{\sigma^2} - 1) * e^{2\mu + \sigma^2}$

When running a linear regression with a log transformed response, each predicted value of $\ln(Y)$ is in natural log scale and should follow a normal distribution, $N(X\hat{\beta}, \hat{\sigma}^2)$, where $\hat{\beta}$ and $\hat{\sigma}^2$ are the estimated regression coefficients and the mean squared error (MSE) from a regression model, respectively. It is important to note that when we exponentiate the predicted value of $\ln(Y)$, we get the predicted geometric mean of Y rather than the predicted arithmetic mean of Y . Using the equation $e^{X\hat{\beta} + \frac{1}{2}\hat{\sigma}^2}$ instead will give the predicted mean values \hat{Y} in the original scale.

Interpreting parameter estimates in a linear regression when variables have been log transformed is not always straightforward either. The standard interpretation of a regression parameter β is that a

¹ The natural- logarithm (denoted by \ln) is used throughout this newsletter.

one-unit change in the predictor results in β units change in the expected value of the response variable while holding all the other predictors constant. Interpreting a log transformed variable can also be done in such a manner. However, such coefficients are routinely interpreted in terms of percent change. Below we will explore the interpretation in a simple linear regression setting when the dependent variable, or the independent variable, or both variables are log transformed. (See the appendix for derivations and formulas)

Consider an example studying the relationship between height and weight. People's weights tend to have a higher variance for taller people, so it is quite reasonable to take log of weight when you are fitting a linear regression model.

Suppose the dependent variable is log-transformed, and the regression is estimated as follows:

$$\ln(\textit{Weight}) = 2.14 + 0.00055 * \textit{Height}$$

The estimated coefficient of the Height variable is $\beta_1 = 0.00055$, so we would say that an increase of one-unit in the *Height* would result in $(e^{\beta_1} - 1) * 100$ percentage change in Y, approximately 0.055% change in the *Weight*.

If the independent variable is log-transformed, the regression equation is:

$$\textit{Weight} = 3.94 + 1.16 * \ln(\textit{Height})$$

Here $\beta_1 = 1.16$. We would say that a one percent change in *Height* is associated with $1.16 * \ln\left(\frac{101}{100}\right) \approx 0.0116$ change in *Weight*.

If both the dependent variable and independent variable are log-transformed, the fitted regression is:

$$\ln(\textit{Weight}) = 1.69 + 0.11 * \ln(\textit{Height})$$

For this model, $\beta_1 = 0.11$. We would conclude that one percentage change in *Height* results in $[(1.01)^{\beta_1} - 1] * 100$ percentage change in Y, or around a 0.11% change in *Weight*.

As always if you would like assistance with this topic or any other statistical consulting question, feel free to contact statistical consultants at CSCU.

Author: Jing Yang (jy543@cornell.edu)

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at <http://www.cscu.cornell.edu>).

Appendix

Interpreting log-transformed parameter estimates in regression models - Formulas

Dependent variable is log transformed

A linear regression model with a log-transformed dependent variable and two predictor variables can be expressed with the following equation:

$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\left. \begin{array}{l} \text{Suppose, } \ln(Y_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ \ln(Y_2) = \beta_0 + \beta_1 (X_1 + 1) + \beta_2 X_2 \end{array} \right\} \beta_1 = \ln\left(\frac{Y_2}{Y_1}\right) \Leftrightarrow e^{\beta_1} = 1 + \frac{Y_2 - Y_1}{Y_1}$$

Then, we have the percent change in $Y = (e^{\beta_1} - 1) * 100$.

The interpretation is that the dependent variable changes by $100(e^{\beta_1} - 1) \approx 100\beta_1$ **percent** for a one unit increase in X_1 while holding all other predictors constant. Here the simple approximation works only when β_1 is small², e.g. less than 0.1.

Independent variable is log transformed

A linear regression model with one log-transformed predictor variable can be expressed with the following equation:

$$Y = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2$$

$$\left. \begin{array}{l} \text{Suppose, } Y_1 = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2 \\ Y_2 = \beta_0 + \beta_1 \ln[X_1 * (1 + 1\%)] + \beta_2 X_2 \end{array} \right\} Y_2 - Y_1 = \beta_1 \ln\left(\frac{1 + 1\%}{1}\right) = \beta_1 \ln\left(\frac{101}{100}\right) \approx \frac{\beta_1}{100}$$

The interpretation is that one **percent** change in the independent variable is associated with $\beta_1 \ln(1.01) \approx \beta_1/100$ change in the dependent variable while all other variables in the model are held constant.

² Here, all of the approximations can only happen using the natural log transformation, not other logarithms.

Both dependent and independent variables are log transformed

When both dependent and independent variables are log transformed, the model can be expressed with the following equation:

$$\ln(Y) = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2$$

$$\left. \begin{array}{l} \text{Suppose, } \ln(Y_1) = \beta_0 + \beta_1 \ln(X_1) + \beta_2 X_2 \\ \ln(Y_2) = \beta_0 + \beta_1 \ln[X_1 * (1 + 1\%)] + \beta_2 X_2 \end{array} \right\} \ln\left(1 + \frac{Y_2 - Y_1}{Y_1}\right) = \beta_1 \ln\left(\frac{101}{100}\right)$$

Then we have the percent change in Y = $[(1.01)^{\beta_1} - 1] * 100$.

So, we can interpret as: one **percentage** change in X results in $100(1.01^{\beta_1} - 1)$ **percentage** change in Y while holding all other variables constant. For β_1 less than 10, $100(1.01^{\beta_1} - 1)$ can be approximated by β_1 .

It's easy to get confused when interpreting percentage change. Here's an example of the correct way to think about it: a change of 80 **percentage** means that the final value is $(1 + 80/100)$ or **1.8 times** the initial value. A change of -30 **percentage** means that the final value is $(1 - 30/100)$ or **0.7 times** the initial value.