Outbreak, Surveillance and Investigation Reports
Field Epidemiology Training Program, Bureau of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625901734-5, Fax: +6625918581, Email: osireditor@osirjournal.net, http://www.osirjournal.net

The Grammar of Science: Let's 'Log' (Part 2)

Jaranit Kaewkungwal*

Mahidol University, Thailand

* Corresponding author, email address: jaranit.kae@mahidol.ac.th

Part 1 of this paper is available at:

Kaewkungwal J. The grammar of science: let's 'log' (Part 1). OSIR. 2017 Jun;10(4):23-27.

Last time, we have shown how 'Log" play roles in mathematics and statistics. Now we will take a close look at how it applies in data management and analysis.

Statisticians also love "log transformed" data

Many statistical procedures have the assumption that the variables in the model should be normally distributed. A significant violation of the assumption can increase errors in study conclusion, depending on the nature of the methods used and the level of non-normality[1]. Even though we can avoid such limitation by using non-parametric statistics that has no explicit assumption about normality, we may sometimes still face with inconclusive results due to the effect of severe non-normally distributed data[2-3].

When our data are not normal, we should explore the reasons behind it. The non-normality may be due to mistakes in data entry (not real extreme-value data), presence of outliers, or the nature of the variable itself. Let's look only at the issue of the latter case where skewedness is due to the nature of variable itself. There are variables in biomedical and clinical study that are almost always not normal, e.g., viral load, titre, length of stay in hospital admission, survival time, etc. But we want to use statistical procedures that require normality assumption for those variables. One way to do it and most commonly used is to do "data transformation" or changing the scale of the data. Data transformation is not cheating, but rather look at data in another way, for example, we can say that 4 is equivalent to square-root of 16 ($\sqrt{16}$). When we change the scale of the data, the distribution will change; generally the extreme values will be pulled closer, e.g., $\sqrt{9}\rightarrow 3$, $\sqrt{16}\rightarrow 4$, $\sqrt{25}\rightarrow 5$. There are many valid reasons for utilizing data transformations, not only for changing the non-normality characteristics but also for improving variance stabilization, conversion of scales to interval measurement, etc.[1-4]

Three data transformations most commonly used in handling non-normality included: square root, logarithm, and inverse. If the distribution of a variable has a positive skew, "log transformation" will usually be used to make that positively skewed distribution to be more approximately normal[4]. As an example, if we plot the histogram of viral load collected from HIV-infected patients, we will see a significant right skew in this data (most patients had low amount of viral loads but a few had extreme amount of viral loads). After we "take log" of the raw data of viral loads, then we plot the histogram of the logarithm of viral loads, we now see a distribution that looks much more like a normal distribution as shown in figure 1.

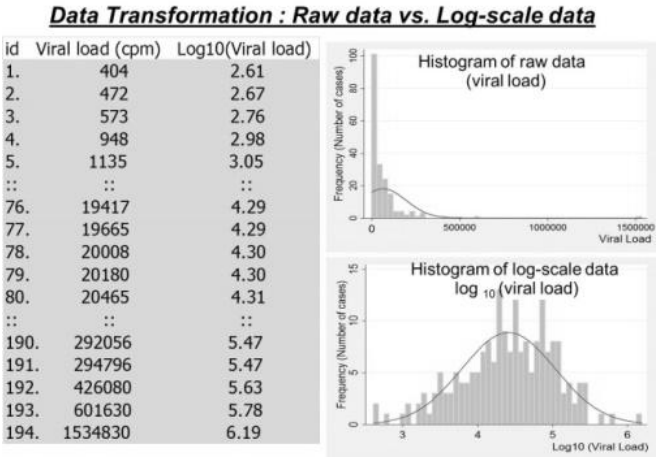

Figure 1. Data transformation in Log base 10 scale (Viral load of HIV patients)

In this example, we took log base 10, but we could take "natural log" or log of other bases and getting somewhat similar normality pattern from different

scaling of data transformation. However, when we interpret the results of the statistical procedures, we have to explain that transformed variable in log-scale, or we have to "anti-logarithm" the results of that variable back to original scale ($\log_{10}X \rightarrow 10^X$, $\ln(X) = \log_e X \rightarrow e^X$, etc)

What is "Logit" in logistic regression?

Before we talk about "logit" in "Logistic regression", let's start with the basic "Linear Regression". Linear regression is a statistical technique for relating the outcome or dependent variable (Y) to one or more predictors or exploratory/independent variables (X). The model is based on a linear relationship between the expected value of Y (y-hat) and each independent variable (when the other independent variables are held fixed)[5].

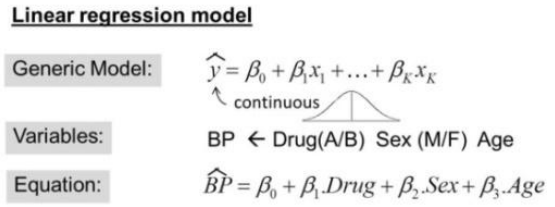

Figure 2. Linear regression model

As shown in figure 2, the "structural model" (generic model) would tell us that when exploratory variable (X) change for 1 unit, the outcome Y would change about β (after controlled for or adjusted for other exploratory variables in the equation). In other words, the structural model describes how the mean response of Y changes with X[5-7]. Based on the example of variables in linear regression equation in figure 2, we can say that the mean differences of blood pressure (BP) between patients taking Drug A vs. Drug B is about β_1; between male vs. female patients, about β_2; and between those ages difference of 1 year, about β_3.

There are several assumptions in fitting the linear regression model. Historically, the normal distribution had a pivotal role in the development of regression analysis and it continues to play an important role[6]. Assumptions about outcome variables are that Y should be normally distributed and variance of Y should be constant[5-8]. When the variance of the Y is not constant, it will lead to violation of another assumption that the error variance in the model becomes not constant (or a fancy term - assumption about homoscedasticity in Y). The assumption about error variance, so-called the "error model", indicates that for each particular X, if we have or could collect many subjects with that x value, their distribution around the population mean should also be normally distributed. The error model suggests that the linear regression not only assumes "normality" and "equal variance", but also the assumption of "fixed-X" (i.e., the explanatory variable is measured without error)[7-8].

When the assumptions are significantly violated, the results of the analysis may be incorrect or misleading. For example, if the assumption of independence of variables in the model is violated, then model may not be appropriate. If the assumption of normality is violated, or outliers are present, then the linear regression goodness of fit test may not be the most powerful or informative test available[5,7].

When we encounter a problem with the equal variance or normality assumptions, we may solve it by using data transformation either using log(y) or y^2 or $\sqrt{y}$ or 1/y instead of y for the outcome. But if we get into non-linearity relationship between exploratory and outcome variables, we may try transformation of X, Y, or both. In fact, this generic model written as "linear" in "linear regression" does not imply that it can apply for only linear relationships. If we transformed X or Y then we could assess non-linear relationships to be represented on a new scale that makes the relationship linear. However, technically the β's must not be in a transformed form[7-8].

Now let's discuss about "logistic regression".

The logistic regression model is a statistical technique for presenting the relation between a binary response or a multinomial response/outcome (Y) and several predictors or exploratory variables (X)[9]. This type of outcome is very common in the field of health science and others, say die - not die, cured - not cured, mild – moderate – severe, etc. Historically, the "logistic function" was originally invented for the purpose of describing the population growth and it was evolving by many statisticians in several academic fields in the US and European. The "logistic regression" name was given by a Belgian mathematician, Pierre François Verhulst (1804-1849)[10].

We could say that the emergence of the logistic function started from the growth curve and mathematically it was evolved making it a close resemblance to the normal distribution function[10,11].To make it easier in explaining the basic concept of logistic regression, let's follow the same idea of linear regression. The model is based on the same generic model of linear relationship between the expected value of outcome Y (y-hat) and each exploratory variable (when the other exploratory variables are held fixed). The difference is that Y in

linear regression is continuous but Y is logistic regression is categorical.

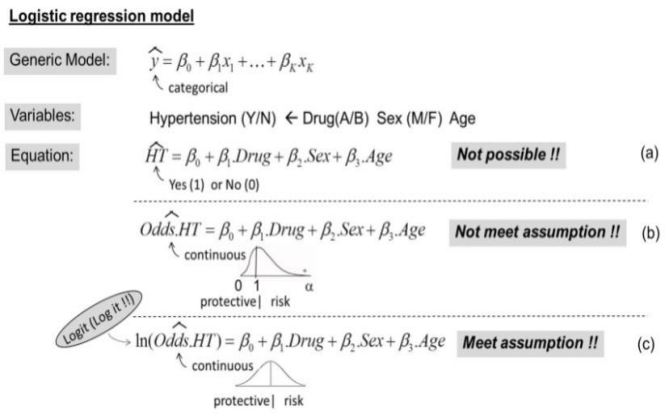

Figure 3. Logistic regression model

P/(1-P)		Odds	Log_{10}(Odds)		Log_e(Odds)
00001/99999	0	0.00001	-∝	-5.0000	-∝	-11.5129
0001/9999		0.00010		-4.0000		-9.2102
001/999		0.00100		-2.9996		-6.9068
01/99		0.01010		-1.9956		-4.5951
1/9		0.11111		-0.9542		-2.1972
50/50	1	1	0	0.0000	0	0.0000
9/1		9		0.9542		2.1972
99/01		99		1.9956		4.5951
999/001		999		2.9996		6.9068
9999/0001		9999		4.0000		9.2102
99999/00001	+∝	99999	+∝	5.0000	+∝	11.5129

Figure 4. Odd and Log (odds)

Figure 3 presents an example of logistic regression model of binary outcome. We now don't want to know the expected value of BP like in linear regression but we want to assess the expected value of hypertension (HT), whether the person has or do not have HT, coding as 1 and 0, respectively. As shown in equation (a), it is not possible mathematically to get expected value of Y (as 0 or 1) from the calculation of known values of Xs and the estimated β's in the equation.

The problems are also about the assumptions of the generic model. As previously mentioned the main assumptions of linear regression are about the "error model" that the errors or residuals (distances of each X around the expected mean of Y) are normally distributed and Y does have to be continuous and measured on an interval or ratio scale[5-8]. Unfortunately, our Y (HT) now is a categorical variable and it could not fit these assumptions. No matter what data transformations, we could not get normal residuals from a model with a categorical response variable[12-14].

Since we cannot use the equation to get the expected value Y of 0/1, we then say that we want to use the equation to explain the "odds" of getting the outcome Y[14-16]. "Odds" is defined as the probability (p) that the event Y occurs (Y=1) over the probability (1-p) that the event Y does not occur (Y=0). "Odds" is (p/1-p). As shown in figure 4, "odds" is now a continuous number, ranging from 0 to infinity. But we still have a problem about the model assumption! Let's look at the concept of "odds". When we have 100 people walk by and 50 of them have the disease (Y occur) and 50 do not have the disease (Y not occur), the odds will be (50/100)/ (50/100) = 1. When we have 10 people walk by and 9 of them have the disease and 1 do not have the disease, the odds will be (9/10)/(1/10) = 9. On the opposite scenario, When we have 10 people walk by and 1 of them have the disease and 9 do not have the disease, the odds will be (1/10) / (9/10) = 0.1111.

Back to the regression model, as shown in equation (b) of figure 3, we now can substitute the values of Xs and the estimated β's to calculate for outcome that is now continuous. But the assumption still does not hold regarding normally distributed of errors and non-linearity of the model. This is because our outcome (odds of Y) is still not normal due the fact that "odds' is positive skewed, ranging from 0-1 for protective side (fewer subjects have the outcome Y) and 1-infinity for risk side (more subjects have the outcome). So equation (b) is not quite appropriate and does meet the basic assumptions.

What can we do? Back to what we discussed before, when linearity fails to hold, even approximately, it may be possible to transform the variables in the regression model to improve the linearity. And if regression on the transformed scale appears to meet the assumptions of linear regression, then we may decide go with the transformations[4-12]. Again, when the data is positively skewed, logarithm is the common way that statisticians use to make the data normally distributed. Regression attempts to model the relationship between exploratory and outcome variables by fitting an equation to observed data. The "logarithm" concept is also about relationship between time and growth. The analogy is that in logarithm we ask "as time change, how much is the growth" and in regression "as an exploratory variable (X) changes, how much is the outcome (Y)". As shown in figure 4, the "odds" after transformed into log scale, either common or natural log would become approximately normally distributed.

The final equation (c) in figure 3 then appears to meet the assumptions. The expected outcome Y is now ln(odds), so-called "logit" term. Thus the logistic regression model is simply a non-linear transformation of the linear regression[13-14].

So we can now tell that when exploratory variable (X) change for 1 unit, the ln(odds) of having the outcome Y would change about β (after controlled for or adjusted for other exploratory variables in the equation). Based on the example of variables in the logistic regression equation in figure 3, we can say that the ln(odds) or ln(p/1-p) of having HT between patients taking Drug A vs. Drug B is about β_1; between male vs. female patients, about β_2; and between those ages difference of 1 year, about β_3.

But how do we tell the patients - if they take Drug A, their ln(odds) to have HT is β_1? No patients will understand that! To make it meaningful – let's simply focus on effect of Drug on odds of getting HT as shown in figure 5. If you take Drug A (code 1), the equation will tell you that ln(odds of HT) = $\beta_0+\beta_1$; but If you take Drug B (code 0), the equation will tell you that ln(odds of HT) = β_0. That means, ln(odds) of the two groups are different by β_1. Solving the equation of subtraction of ln(odds) of the two groups, we get division in log scale (conversion rules between division and subtraction!). The odds of group 1 (Drug A) vs. odds of group 0 (Drug B) is called "odds ratio" (OR). This OR will tell us how much the two groups are different in terms of chance to get HT over chance of not getting HT.

But still based on solving the equation (a)-(b) as shown in figure 5, we do not yet have OR, but have ln(OR) =β_1. No patients will understand that ln(OR)! In most cases, when we report the result, we have to "back transform" the expected value (point estimates and its confidence intervals) from the model for better interpretability. The "back transform" is the inverse of the transformation to return to the original scale; that is, the antilogarithm. In case of this logistic regression model, the inversion of the equation, ln(OR) =β_1, becomes OR = $e^{\beta 1}$. Thus, after we estimate β_1 by fitting the logistic regression model, we can then simply exponential it. And we now can explain to our patients how much the two groups are different in terms of their odds of having the outcome!

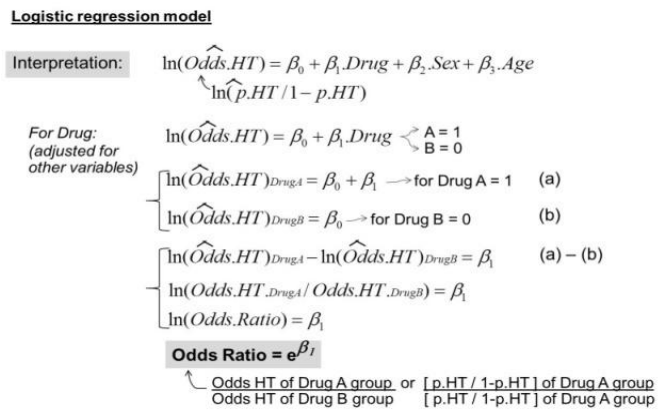

Figure 5. Interpretation of logistic regression model

Beyond "Log"

Logarithm is used a lot more in different statistical techniques. Some make argument on the limitation of "logarithm" that it cannot handle negative numbers. But Euler had once said "To those who ask what the infinitely small quantity in mathematics is, we answer that it is actually zero. Hence there are not so many mysteries hidden in this concept as they are usually believed to be."

Since natural logarithm is used quite often to explain relationship of changes, I would like to end this "Let's Log" with Euler's equation that is considered as the "beautiful equation"[17] of all and proved to be true, $e^{i\pi}$ - 1 = 0. Interestingly, 1 and 0 are real numbers, e and π are irrational numbers (values that can't be given precisely in decimal notation) and i is the "imaginary" number which is √ -1 (mathematically invented imaginary number as doubling -1 can never get -1). An imaginary number seems strange but getting real number from the power (inverse of logarithm) of an imaginary number and irrational numbers is even awesome (rockin!).

Suggested Citation

Kaewkungwal J. The grammar of science: let's 'log' (Part 2). OSIR. 2017 Sep;10(3):22-26.

References

1. Osborne JW. Notes on the use of data transformations. Practical Assessment, Research & Evaluation. 2002 May;8(6):2002.
2. Zimmerman DW. Increasing the power of nonparametric tests by detecting and down weighting outliers. Journal of Experimental Education. 1995;64(1):71-8.
3. Zimmerman DW. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. Journal of Experimental Education, 1998;67(1):55-68.
4. Benoit K. Linear regression models with logarithmic transformations. 2011 Mar 17 [cited 2017 Apr 4]. <http://kenbenoit.net/assets/courses/ME104/logmodels2.pdf>.
5. Feinberg School of Medicine, Northwestern University. PROPHET Stat Guide: Do your data violate linear regression assumptions? 1997 Mar 14 [cited 2017 Apr 4]. <http://www.basic.northwestern.edu/statguidefiles/linreg_ass_viol.html>.

6. Frees EW. Regression modeling with actuarial and financial applications. Cambridge: Cambridge University Press; 2010.
7. Seltman HJ. Experimental design and analysis. 2009 [cited 2017 Apr 4]. <http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>.
8. Department of Statistics, Yale University. Inference in linear regression [cited 2017 Apr 4]. <http://www.stat.yale.edu/Courses/1997-98/101/linregin.htm>.
9. Menard SW. Applied logistic regression analysis. California: Sage; 2002.
10. Wilson JR. & Lorenz KA. Modeling binary correlated responses using SAS SPSS and R. Switzerland: Springer International Publishing; 2015.
11. Analytics Vidhya Content Team. Going deeper into regression analysis with assumptions, plots & solutions. 2016 Jul 14. [cited 2017 Apr 4]. <https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/>.
12. Grace-Martin K. What is a Logit function and why use logistic regression? [cited 2017 Apr 4]. <http://www.theanalysisfactor.com/what-is-logit-function/>.
13. Appalachian State University, North Carolina. An introduction to logistic regression, nuts and bolts [cited 2017 Apr 4]. <http://www.appstate.edu/~whiteheadjc/service/logit/intro.htm>.
14. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb 15;24(1):12-8. eCollection 2014.
15. McDonald JH. Handbook of biological statistics. 3rd ed. Maryland: Sparky House Publishing; 2014.
16. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley & Sons, Inc.; 2000.
17. Rehmeyer J. Euler's beautiful equation. Science News. 2007 Apr 12 [cited 2017 Apr 4]. <https://www.sciencenews.org/article/eulers-beautiful-equation>.