Use of Ratios and Logarithms in Statistical Regression Models

Scott S. Emerson, M.D., Ph.D.

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

January 22, 2014

Abstract

In many regression models, we use logarithmic transformations of either the regression summary measure (a log link), the regression response variable (e.g., when analyzing geometric means), or one or more of the predictors. In this manuscript, I discuss the rationale for using logarithmic transformations, the interpretation of ratios, and the general properties of logarithms.

1 Use of Logarithmic Transformations

Logarithmic transformations of data and/or parameters are used extensively in statistics. The fundamental reason for this stems from the following logic:

1. We are most often interested in using statistics to detect associations between two variables.
2. By an association, we mean that the distribution of one variable (we call this the "response variable") is different in some way between groups that are homogeneous for the other variable (we call this the "predictor of interest" or POI).
3. If the two variables are associated, then that means that some aspect of the distribution (some "summary measure" like the mean, geometric mean, etc.) is unequal between two groups that differ in their value of the POI. In describing the association, we will want to describe
   (a) How the POI differs between the two (conceptual) groups being compared, and
   (b) How the summary measure of response compares across two groups.
4. There are two simple arithmetic ways to tell if two numbers (e.g., the level of POI in two separate groups, or the mean of the "response variable" in two separate groups) are unequal:
   (a) their difference is not 0, or
   (b) their ratio is not 1.
5. We choose between differences and ratios as methods for comparisons based on a variety of criteria that are sometimes competing:
   (a) People understand differences more than they understand ratios.
      i. Part of this is because natural language (e.g., English) is not very precise when describing ratios.

ii. But I may have the cause and effect relationship reversed: Natural language is bad at describing ratios, because the average speaker (i.e., human) does not understand them, and thus humans did not invent natural language sufficiently precise to describe ratios.

(b) Differences are better at describing the scientific importance of most comparisons.

i. For instance, you probably care less about having 10 times as much money as me if I only have one cent (a difference of $0.09).

ii. You probably care more about having $1,000,000 more than me, even if I have $10,000,000 (a ratio of only 1.1).

(c) When working with very small numbers, however, a ratio will accentuate an effect better than a difference will.

i. Being diagnosed with lung cancer in any given year is a very rare event in nearly every population, including smokers. (In the numbers that follow, I am combining data from many different sources. The order of magnitude is probably correct, but the exact numbers may be a bit off.)

ii. In the US, 60-64 year old current or former smokers have a probability of 0.00296 to be diagnosed with lung cancer during the next year.

iii. In the US, 60-64 year old never smokers have a probability of 0.000148 to be diagnosed with lung cancer during the next year.

iv. The difference in cancer incidence rates is thus a very small 0.002812.

v. However, smokers have about a 20-fold higher rate of cancer diagnosis than non-smokers of the same age and sex. (It actually seems to vary by sex, age, race, but the point still obtains.)

(d) Sometimes scientific mechanisms dictate that ratios are more generalizable for the summary measure of the response distribution and/or for effects due to predictors.

i. Interventions and risk factors often affect the rate that something happens over time.

ii. Cellular enzymes affect the rate at which biochemical reactions proceed. Risk factors that affect enzymatic activity will change the amount of some chemical that accumulates.

A. Many physiologic actions occur when some agonist A (e.g., a chemical or drug) interacts with a receptor R (e.g., a particular region of an enzyme or portion of a cell membrane). The activity occurs when a receptor-agonist complex RA is formed.

B. A simple (simplistic) model for this interaction is given by Michaelis-Menten kinetics, where the relative abundance of the receptor-agonist complex is governed by some reaction constant K_{eq} that relates the concentration $[RA]$ of the receptor-agonist to the product of the concentrations of the free receptor and agonist ($[R]$ and $[A]$):

$$R + A \rightleftharpoons RA \qquad \Longrightarrow \qquad K_{eq} = \frac{[RA]}{[R]\,[A]} \qquad \Longrightarrow \qquad \frac{[RA]}{[R]} = K_{eq}[A]$$

C. Clinical outcome measures are often most directly related to the relative abundance of the receptor-agonist complex, and from the above equation we see that drugs or risk factors that affect the rate of reaction through the K_{eq} act multiplicatively, rather than additively, on the concentration of the agonist.

iii. The actions of many biochemical pathways are influenced heavily by the rates of absorption and excretion. Quite often these rates are proportional to the concentration of the drug. Hence the biochemical concentration C_t at any point in time t follows an exponential decay model

$$C_t = C_0 e^{-Kt}$$

and drugs that change the rate parameter K act multiplicatively, rather than additively, on the initial concentration C_0.

iv. When dealing with money (e.g., health care costs), we most often apply taxes and interest as a rate (percentage) and we (therefore) measure inflation on a multiplicative scale.

A. Hence, the fact that women's starting salaries are often lower than men's starting salaries by some amount leads to even greater differences after a number of years, though the ratio will tend to be constant, because ratios are given as a percentage increase each year.

B. Similarly, compared to the prices when I graduated from high school, costs in 2013 are approximately 5-fold higher across a wide variety of products, though the difference in the cost of movies from 1973 to 2013 (about $8.00) and the difference in the cost of cars (that I would buy) from 1973 to 2013 (about $18,000) are not at all the same.

(e) Taking differences is generally easiest, and it is generally most stable statistically, because denominators that tend toward zero cause wild fluctuations in ratios.

(f) And there are some highly technical reasons: In certain distributions, the logarithmic transformation of some common distributional summary measure can be shown to be efficiently estimated using unweighted combinations of the observations (so every subject can be treated equally). That is, for those specific distributions models based on the log link function will in some sense be "nicer".

i. For a Bernoulli random variable (a variable that is binary), the log odds (logit mean) is the "canonical parameter".

ii. For a Poisson random variable, the log rate (log mean) is the "canonical parameter".

iii. For a log normal random variable (with known σ^2), the log geometric mean is the "canonical parameter".

6. When ratios are scientifically or statistically preferred, we gain stability by considering the logarithm of the ratios, because (as will be demonstrated in later sections of this document) the logarithm of a ratio is the difference between the logarithm of the numerator and the logarithm of the denominator. Hence, by using logarithms, we are back on an additive scale.

Note: In the above motivation for the use of logarithms, noticeably absent is any reference to transformations to obtain normal distributions. This is not a reason to transform data. I note that it is easy to demonstrate times that non normally distributed data (either as response or predictors) are more efficiently analyzed in their untransformed state than when transformed to a normal distribution. Instead, we do like to work on scales where effects of covariates are additive, rather than multiplicative. But it is true that skewed data often arise through mechanisms described above, so many data analysts (and authors) have gotten "the cart before the horse" and regarded that skewness the reason for log transformation. (For instance, we can have a truly linear relationship between untransformed variable Y and untransformed variable X, but I may have sample X with some large outliers, and thus the distribution of Y is similarly skewed. In this setting, there may be some very influential points, but under the conditions I laid out, there is no justification for transforming either variable.)

2 Examples of Variables that are Often Logarithmically Transformed

A number of commonly encountered scientific quantities are so typically used after logarithmic transformations, that the measurements themselves are almost always expressed on a logarithmic scale. Examples include:

- Acidity / alkalinity of an aqueous solution is measured as the hydrogen ion concentration. However, it is most common to report the pH, which is the negative log base 10 transformation of the hydrogen ion concentration. A 1 unit change in the pH thus corresponds to a 10-fold increase or decrease in the hydrogen ion concentration.

- In acoustics, the sound pressure is typically measured on a multiplicative scale. Hence, we consider the sound pressure relative to a standard pressure. That standard pressure is typically based on human hearing in the medium (a different standard is used for air versus water). We refer to that ratio on the logarithmic (base 10) scale using the unit “bel”, or more commonly as a value that is 10 times the $\log_{10}$ scale as a “decibel” of ”dB”. A 3 dB increase in sound thus represents an approximate doubling of the sound pressure, because $\log_{10} 2 = 0.3010$ and $10 \times .3010 = 3$.

- In seismology, the strength of earthquakes is quantified by the “moment magnitude scale (MMS)”, which is a successor to the Richter scale. It is two-thirds the logarithmic (base 10) transformation of the seismic moment, minus 10.7. Because the base 10 logarithm is multiplied by 2/3, a 1,000 fold increase in the strength of an earthquake is measured as a 2 unit difference on the MMS.

In biomedicine, it is very common to use logarithmic transformations for measurements of antibody concentration and mRNA concentration (gene expression), because these concentrations differ by orders of magnitude across individuals (and sometimes within individuals over time). Similarly owing to the exponential growth associated with viral replication, viral load in hepatitis C or HIV research is generally analyzed after logarithmic transformation.

For other measures of concentration, our habits will differ according to the populations considered.

Physiologic homeostasis (regulation to maintain balance or equilibrium in a state conducive to healthy life) tends to maintain important components of the blood in relatively tight control in healthy people. Hence, even though we measure concentrations (and the K_{eq} of various chemical reactions would argue that multiplicative effects would still be relevant), the levels are so constant that the log function is relatively constant over the range of observed data. In health, then, it is relatively unimportant that we log transform the measures. For instance, over the “normal” range of serum bilirubin (0.3 - 1.1 mg/dL, or so depending on the laboratory), the following graph displays the natural log ($\log_e$) of bilirubin versus bilirubin. A straight line, while not perfect, is still a good approximation to the curve.

However, homeostasis is deranged in disease, and levels of specific blood components are uncontrolled. In that setting, the log transformation will be more important to capture the truly important differences in levels. Continuing the example using serum bilirubin, the following graph displays the log of bilirubin versus bilirubin in a population of 418 Mayo Clinic patients who have primary biliary cirrhosis (www.emersonstatistics.com/Datasets/liver.doc). When the pathologic, extremely high measurements of bilirubin are included (the maximum bilirubin in this dataset was 28 mg/dL), there is marked departure from a straight line.

The issue then is whether it is likely that the serum bilirubin is linear in its ability to predict severity of disease. That is, if an elevated bilirubin of, say, 2.0 mg/dL is suggestive of more advanced disease and therefore increased risk of death relative to a PBC patient with a “normal” bilirubin of 1.0 mg/dL, would we really expect any such increased risk to be linear over a range that extends to 28 mg/dL? I would argue no for two reasons, one pathophysiologic and one empirical.

First, as noted above, many physiologic mechanisms act on a multiplicative scale by virtue of the chemical reactions associated with absorption and excretion kinetics. Primary biliary cirrhosis is a disease of unknown etiology that affects the ability of the body to excrete bilirubin. Hence, we expect the diseased state to accumulate bilirubin on a multiplicative scale: Each “step” in disease progression should result in a multiplicative increase in bilirubin. But it is not the bilirubin that is harmful *per se* (at least not in adults–in children the worst effects of kernicterus are directly related to high levels of serum bilirubin being deposited in the developing brain). Instead, serum bilirubin is just a marker of more advanced disease, and we would want to use a measure that is more closely aligned with stage of disease.

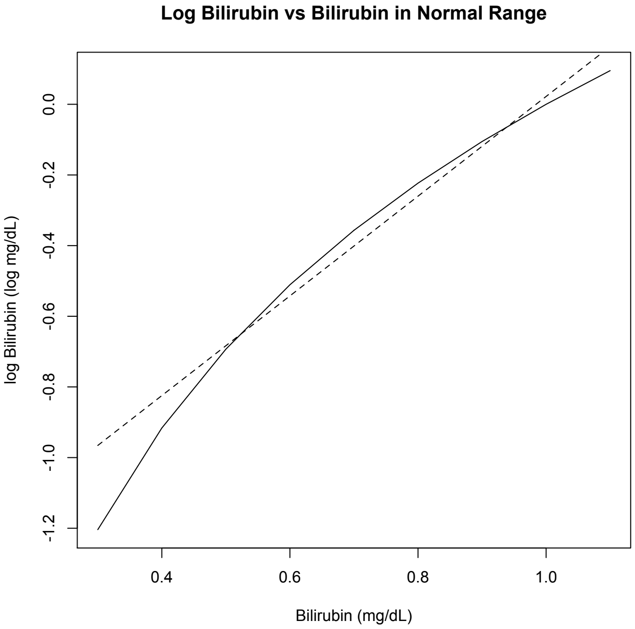

Figure 2.1
A plot of the logarithmic transformation ($\log_e$) of serum bilirubin versus serum bilirubin for the 177 patients in the Mayo primary biliary cirrhosis data set who have serum bilirubin within a normal range of 0.3 to 1.1 mg/dL.

Empirically, we can consider the likelihood that the association between disease outcome of interest (in this instance, death over an observation period that extended up to 13 years) and serum bilirubin (as a marker of disease stage) would be additive or multiplicative over the range of observed values. It is known from clinical experience and many prior studies that treated PBC patients with bilirubin of 2 - 3 mg/dL are at increased risk of death relative to those whose treated bilirubin levels are in the normal range. This increased risk of death has been estimated to be about a 2-fold increase in the rate (hazard) of death between subjects having a bilirubin of about 2 mg/dL and subjects having a bilirubin of about 1 mg/dL. If an additive scale obtains, then someone with a serum bilirubin of 28 would be expected to have a 2-fold increase in risk of death for each 1 mg/dL *difference* in serum bilirubin. A difference of 27 mg/dL would thus be associated with a $2^{27} = 134,217,728$-fold higher risk of death at any given time. Now, again from previous clinical experience and scientific studies, treated PBC patients with normal serum bilirubin levels have a death rate of about 2.5 deaths per 100 person-years. If that risk were increased by a factor of 1.34×10^8, the patient with a bilirubin of 28 mg/dL should die in front of our eyes. Hence, empirically, our prior evidence about increased risk of death for patients with mild elevations of serum bilirubin and our observation that some patients have extremely elevated serum bilirubin levels tells us that it is highly unlikely that serum bilirubin is a marker of increased risk of death that is accurate on a linear scale.

On a multiplicative scale, our empiric evidence is much more believable. The approximate 2-fold increase in risk of death associated with a treated serum bilirubin of 2 mg/dL compared to 1 mg/dL could also be viewed as a 2-fold increased risk for a *doubling* of serum creatinine. An observed serum bilirubin of 32 mg/dL (close enough to 28 mg/dL for the purposes of this simple exposition) represents five doublings. So on a multiplicative scale, we might expect the risk to be $2^5 = 32$-fold higher, or a death rate of about 78.7 deaths per 100 person-years. *(Technical note: In my "back of the envelope" calculations, I am using analyses appropriate for survival times that follow an exponential distribution. This assumption is unrealistic*

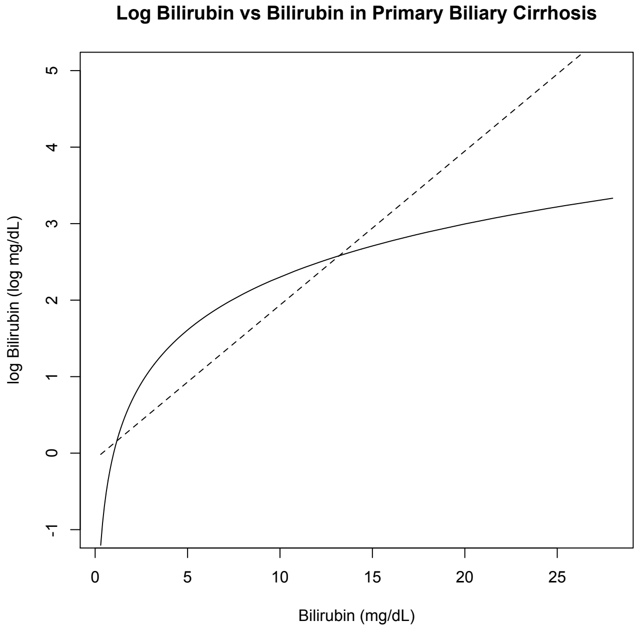

Figure 2.2
A plot of the logarithmic transformation of serum bilirubin versus serum bilirubin for all 418 patients in the Mayo primary biliary cirrhosis data set (range 0.3 - 28 mg/dL).

for human survival over a long period of time, but over shorter periods of time and for diseased populations, it sometimes does not do so badly.)

Other measurements that are commonly log transformed in diseased populations include

- Serum creatinine in kidney disease
- C-reactive protein in the presence of a population with inflammatory components to their disease
- Prothrombin time in patients with clotting abnormalities
- Prostate specific antigen in patients with prostate cancer
- Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), alkaline phosphatase in liver disease
- Antibody titers in autoimmune diseases

3 Examples of Summary Measures that are Often Logarithmically Modeled

In our general regression model, contrasts of some summary measure θ are made across groups defined by covariates $\vec{X} = (X_0 = 1, X_1, X_2, \ldots X_p)$ using the expression

$$g(\theta) = \vec{X}^T \vec{\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

In this expression, we call $\eta = \vec{X}^T\vec{\beta}$ the "linear predictor" that represents the combined "effect" of all the covariates on the value of θ. For a population having $\vec{X} = \vec{x}_i$ we can define the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}.$$

Note that the for the ith population having $\vec{X} = \vec{x}_i$ and the jth population having $\vec{X} = \vec{x}_j$ we find that the difference in linear predictors across the two populations is

$$\eta_i - \eta_j = \sum_{\ell=1}^{p} \beta_\ell (x_{\ell i} - x_{\ell j}) = \beta_1(x_{1i} - x_{1j}) + \beta_2(x_{2i} - x_{2j}) + \cdots \beta_p(x_{pi} - x_{pj}).$$

We term $g(\,)$ the "link function" that links the linear predictor back to the distributional summary measure θ.

A regression model that uses the identity function $g(\theta) = \theta$ is called "additive" on the linear predictor, because differences $(\eta_i - \eta_j)$ in the linear predictors for two populations relate to the *difference* between the values of θ for the two populations: $\theta_i - \theta_j = (\eta_i - \eta_j)$. The commonly used regression model with an identity link is:

- Linear regression: θ is the *mean* of some response variable Y, and $g(\theta) = \theta$ yielding a regression model

$$\theta_{\vec{x}} = E[Y|\vec{X} = \vec{x}] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

A regression model that uses the logarithmic function $g(\theta) = \log(\theta)$ is called "multiplicative" on the linear predictor, because differences $(\eta_i - \eta_j)$ in the linear predictors for two populations relate to the *ratio* between the values of θ for the two populations: $\theta_i/\theta_j = e^{(\eta_i - \eta_j)}$. Invariably, the log link is defined using the natural logarithm $\log_e$. The following commonly used regression models use a log link:

- Logistic regression: $\theta = p/(1-p)$ is the *odds* of some Bernoulli (binary) response variable $Y \sim \mathcal{B}(1,p)$, and $g(\theta) = \log(\theta)$ yielding a regression model

$$\log(\theta_{\vec{x}}) = \log\left[\frac{p_{\vec{x}}}{(1-p_{\vec{x}})}\right] = \log\left[\frac{E[Y|\vec{X} = \vec{x}]}{(1 - E[Y|\vec{X} = \vec{x}])}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- Poisson regression: θ is the *mean* of some positive response variable Y, and $g(\theta) = \log(\theta)$ yielding a regression model

$$\log(\theta_{\vec{x}}) = \log\left[E[Y|\vec{X} = \vec{x}]\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- Proportional hazards regression: $\theta = \lambda(t)$ is the *hazard function* of some response variable Y, and $g(\theta) = \log(\theta)$ yielding a regression model

$$\log(\theta_{\vec{x}}) = \log\left[\lambda_Y(t|\vec{X} = \vec{x})\right] = \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Note: The definition of θ and $g(\,)$ in the above regression models is my preferred interpretation. But it should be noted that other interpretations are possible. For instance, in Poisson regression, we could say that $\theta_{\vec{x}} = \log(E[Y|\vec{X} = \vec{x}])$ and that we use the identity link $g(\theta) = \theta$. When the "generalized linear model" was defined, it was always considered that θ was the mean of the response variable. In that setting then, logistic regression is interpreted as a model of $\theta_{\vec{x}} = E[Y|\vec{X} = \vec{x}]$ with logit link $g(\theta) = \log[\theta/(1-\theta)]$ yielding the exact same regression model:

$$\text{logit}(\theta_{\vec{x}}) = \log\left[\frac{p_{\vec{x}}}{(1-p_{\vec{x}})}\right] = \log\left[\frac{E[Y|\vec{X} = \vec{x}]}{(1 - E[Y|\vec{X} = \vec{x}])}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

The only thing that changes is the interpretation of the regression parameters in terms of θ and vice versa.

4 Review of Logarithms

1. Recall from basic algebra that when you multiply two numbers, you add exponents. That is, if you want to multiply 10^3 time 10^7, the answer is 10^{10}.

2. This only works when you express the numbers as exponents of the same base. Hence, we can not so easily multiply 2^3 times 4^5. Instead, we would want to convert each number to be a power of the same base. In the problem I have given here, this is easy, because $4 = 2^2$. Hence, $4^5 = (2^2)^5 = 2^{10}$.

3. It is possible to raise a base number to a fractional power. For instance, $4^{0.5}$ is just the square root of 4, or 2. Similarly, $81^{0.25}$ is the fourth root of 81 (the square root of the square root), or 3.

4. Before calculators were in widespread use (I can remember back that far), logarithms were used to make multiplication problems easier. That is, every number was converted to an exponential form, the exponents were added, and then the answer was converted back.

5. In this process, some common base for the exponential form would have to be chosen. Commonly that base was 10 for tables of logarithms. The logarithm base 10 of a number was just the exponent of the number expressed as a power of 10. For instance, because $10^2 = 100$, the logarithm base 10 of 100 is 2. Similarly, the logarithm base 10 of 1000 is 3, because $10^3 = 1000$.

6. Every positive number can be expressed as a power of 10. For instance, $10^{0.3010} = 2$. Finding the appropriate exponent for such a representation (that exponent is termed the logarithm base 10, so the logarithm base 10 of 2 is 0.3010) involves a complicated formula, and in the old days tables were used. Now most calculators have a button you can push to find the logarithm base 10 of a number.

7. More generally, we can talk about the logarithm base k of a number x, which we will write as $\log_k(x)$. k can be any positive number; it does not need to be an integer. If $\log_k(x) = y$, then $k^y = x$. We sometimes speak of the antilog base k of y as being x.

8. In earlier math courses, you probably learned a convention that writing 'log' was understood to mean the base 10 logarithm and writing 'ln' was the natural logarithm (base $e = 2.7182818\ldots$). Like all simple rules, however, this is violated regularly. In fact, it is common in science to use 'log' (with no subscript) to mean the natural logarithm. Many statistical software packages use this convention. We will see that this need not be so much of a problem, however.

9. Using different bases for logarithms is just like measuring length in different units (inches, feet, centimeters, miles, light years). No matter what base you use

$$\log(1) = 0$$

   This is because $k^0 = 1$ for all numbers k.

10. There is a constant of conversion between $\log_e(x)$ and $\log_k(x)$ for any base k. For instance, in the following table of selected base 2, base 10, and base e logarithms

x	$\log_2(x)$	$\log_{10}(x)$	$\log_e(x) = \ln(x)$
1	0.000000	0.0000000	0.0000000
2	1.000000	0.3010300	0.6931472
3	1.584963	0.4771213	1.0986123
5	2.321928	0.6989700	1.6094379
10	3.321928	1.0000000	2.3025851
20	4.321928	1.3010300	2.9957323

You can get every number in one column by multiplying the number in another column by some constant. For instance, every number in the $\log_{10}(x)$ column is just .3010300 times the number in the $\log_2(x)$ column. Similarly, every number in the $\log_e(x)$ column is just 2.3025851 times the number in the $\log_{10}(x)$ column. In general, then, we can find the base k logarithm of any number by either of the following formulas

$$\log_k(x) = \log_{10}(x)/\log_{10}(k)$$
$$\log_k(x) = \log_e(x)/\log_e(k)$$

I know of no statistical packages that do not provide $\log_e(x)$, and most provide $\log_{10}(x)$ as well.

11. Important properties of the logarithm come from the properties of exponents:
    (a) $\log_k(xy) = \log_k(x) + \log_k(y)$
    (b) $\log_k x - \log_k(y) = \log_k(x/y)$
    (c) $\log_k(x^y) = y * \log_k(x)$
12. In this class, as in most of science, we will use $\log(x) = y$ to mean $\log_e(x) = y$. This agrees with the nomenclature of both Stata and R. This also is the base that is used as the link function in regression models, so the antilog of y will be take $e^y = x$.

5 Log Transformations in One- and Two-Sample Problems

In one and two sample problems, there is no reason to transform the predictor of interest (POI):

- In one-sample problems, the POI is constant and generally does not enter into the analysis in any way.
- In two-sample problems, the POI is a binary variable. I usually encourage that POI to be coded as 0 or 1, though it does not truly affect the analysis results returned by standard statistical software for two-sample problems. (It does, however, change the results when two-sample problems are implemented in a regression model.)

Hence, the only issue of concern is how to interpret a statistical analysis when the response variable is transformed.

Suppose we have random variables X_i and Y_i. If we take logarithmic transformations $W_i = \log_e(X_i)$ and $Z_i = \log_e(Y_i)$, then $\overline{W}$ is the natural log of the geometric mean of X, and $\overline{Z}$ is the natural log of the geometric mean of Y. It follows, then, that $e^{\overline{W}}$ and $e^{\overline{Z}}$, are respectively the geometric means of X and Y.

Furthermore, $\overline{W} - \overline{Z}$ is the natural log of the ratio of geometric means. (The log of a ratio is the difference of the logs.) Thus, when we do inference using $\overline{W}$ and $\overline{Z}$, we can easily back transform the data to get the geometric means and ratios of geometric means. Such back transformation works for point estimates and confidence intervals. For instance, $e^{\overline{W}-\overline{Z}} = e^{\overline{W}}/e^{\overline{Z}}$ is the ratio of the geometric mean for X to the geometric mean for Y.

I note that if the log transformed data are symmetric, then the geometric mean and the median are the same number. In that case, we could refer to the ratio of medians. As a general rule, however, a larger sample size is required to be sure that a distribution is symmetric than is required to estimate the geometric means. Hence, I do not really recommend that you presume symmetry. It is safer to just talk about the geometric means.

6 Log Transformations in Linear Regression Models

We will first consider the linear regression model, because in that model we can consider transformations of both the response and predictor variables.

6.1 Untransformed Predictors

Suppose we model

$$E[Y|X = x] = \beta_0 + \beta_1 \times x$$

1. From our standard interpretation of regression slope parameters, we know that every 1 unit difference in X is associated with a β_1 unit difference in the expected value of Y:

   $$E[Y|X = a + 1] - E[Y|X = a] = (\beta_0 + \beta_1 \times (a + 1)) - (\beta_0 + \beta_1 \times a) = \beta_1.$$

   If a straight line relationship holds, this is exactly true for every choice of a. If the true relationship is nonlinear, then β_1 represents some sort of average difference.

2. Similarly, we know that every c unit difference in X is associated with a $c\beta_1$ unit difference in the expected value of Y:

   $$E[Y|X = a + c] - E[Y|X = a] = (\beta_0 + \beta_1 \times (a + c)) - (\beta_0 + \beta_1 \times a) = c\beta_1.$$

6.2 Transformations of Predictors

Suppose we model

$$E[Y|X = x] = \beta_0 + \beta_1 \times \log_k(x)$$

1. From our standard interpretation of regression slope parameters, we know that every 1 unit difference in $\log_k(X)$ is associated with a β_1 unit difference in the expected value of Y.

2. Similarly, we know that every c unit difference in $\log_k(X)$ is associated with a $c\beta_1$ unit difference in the expected value of Y.

3. Now, a 1 unit difference in $\log_k(X)$ corresponds to a k-fold increase in X, and a c unit difference in $\log_k(X)$ corresponds to a k^c-fold increase in X.

   - Ex: A 1 unit change in $\log_{10}(CHOLEST)$ corresponds to a 10 fold increase in $CHOLEST$. A 3 unit change in $\log_2(CHOLEST)$ corresponds to a $2^3 = 8$ fold increase in cholesterol.

4. If we want to talk about a 10% increase in X, then that would correspond to a $c = log_k(1.1)$ unit increase in $\log_k(X)$.

   - Ex: Suppose we model predictor $HEIGHT$ on a log base 10 scale. Because we never see a 10 fold increase in height, when interpreting our model parameters it might be better to consider comparisons between populations which differ in height by, say, 10%. We would then estimate the difference in the expected response as $\log_{10}(1.1)\hat{\beta}_1$, where $\hat{\beta}_1$ was the least squares estimate for the slope parameter in the regression. Note that we would find a confidence interval for the effect associated with that 10% change in height by multiplying the CI for β_1 by $\log_{10}(1.1)$ as well. (If you wanted to get a statistical package to do all this for you, just use the base 1.1 logarithm for height in the regression model:

     $$htlog = \log_e(ht)/\log_e(1.1).$$

     Then a 1 unit change in your predictor corresponds to a 10% change in height.)

6.3 Transformation of Response

Suppose we model (for arbitrary base j)

$$E[\log_j(Y)|X = x] = \beta_0 + \beta_1 \times x$$

1. Using the standard interpretation of regression slope parameters, we know that every 1 unit difference in X is associated with a β_1 unit difference in the expected value of $\log_j(Y)$, and every c unit difference in X is associated with a $c\beta_1$ unit difference in the expected value of $\log_j(Y)$.

2. Unfortunately, a β_1 unit difference in the expected value of $\log_j(Y)$ does not have an easy interpretation in the expected value of Y. However, statements made about the distribution of $\log_j(Y)$ are generally not well understood by the general population, so we need to find another way.

3. The expected value of $\log_j(Y)$ is the log of the geometric mean of Y. Thus, we can make statements about the geometric mean of Y considering our model to be
$$E[\log_j(Y)|X = x] = \log_j(GeomMn[Y|X = x]) = \beta_0 + \beta_1 \times x$$

4. Under this modification, a β_1 unit difference in the base j logarithm of the geometric mean of Y corresponds to a j^{β_1}-fold change in the geometric mean of Y. Similarly, a $c\beta_1$ unit difference in the base j logarithm of the geometric mean of Y corresponds to a $j^{c\beta_1}$-fold change in the geometric mean of Y. We can say that $j^{c\beta_1}$ is the ratio of geometric means for two populations which differ by c units in their values for X.

5. It is probably easiest to use $j = 10$ or $j = e$, because most calculators have a button that will compute the antilogs for those bases.

6. *(A very special case in which we can talk about medians. I truly recommend talking about geometric means, instead.)* I note that under standard classical assumptions of linear regression (which classical assumptions assume normality of residuals), the expected value of $\log_j(Y)$ is also the median of $\log_j(Y)$. (Actually, we do not need normality, but we do need the error distribution to be symmetric about its mean. If you do assume normality, then we can state our assumption as being that Y has the lognormal distribution in each subpopulation.) Thus, we can make statements about the median of Y considering our model to be
$$mdn[\log_j(Y)|X = x] = \log_j(mdn[Y|X = x]) = \beta_0 + \beta_1 \times x$$
Under this modification, a β_1 unit difference in the base j logarithm of the median of Y corresponds to a j^{β_1}-fold change in the median of Y. Similarly, a $c\beta_1$ unit difference in the base j logarithm of the median of Y corresponds to a $j^{c\beta_1}$-fold change in the median of Y. We can say that $j^{c\beta_1}$ is the ratio of medians for two populations which differ by c units in their values for X.

6.4 Transformations of the Response and Predictor

This is just a combination of the above settings. That is, we talk about the ratio of geometric means of Y associated with a several-fold increase in X. Suppose we model (for arbitrary bases j and k)

$$E[\log_j(Y)|X = x] = \beta_0 + \beta_1 \times \log_k(x)$$

1. An r-fold change in X (so a $c = \log_k(r)$ unit difference in $\log_k(X)$) will be associated with an $r^{\beta_1/\log_j k}$-fold change in the geometric mean of Y. That is, the geometric mean ratio of Y is $r^{\beta_1/\log_j k}$ when comparing two populations, one of which has X r times higher than the other.

2. The above formula becomes much easier if the same base is used for both predictor and response. In this case, $j = k$, and the geometric mean ratio is simply r^{β_1} when comparing two populations, one of which has X r times higher than the other.

7 Log Transformations in Regression Models using Log Links

In logistic, Poisson, and proportional hazards regression we use a log link, and those logarithms are invariably on the natural log scale ($\log_e$). Hence we have to consider a different interpretation of the parameters, and in the remaining parts of this document I adopt the standard notation that $\log(x) = \log_e(x)$ and any other base would be explicitly specified (e.g., the base 10 logarithmic function would be written $\log_{10}(x)$.

7.1 Untransformed Predictors

Suppose we model

$$\log(\theta_x) = \beta_0 + \beta_1 \times x$$

1. From our standard interpretation of regression slope parameters, we know that every 1 unit difference in X is associated with a β_1 unit difference in $\log(\theta_x)$:

$$\log(\theta_{a+1}) - \log(\theta_a) = (\beta_0 + \beta_1 \times (a+1)) - (\beta_0 + \beta_1 \times a) = \beta_1.$$

   If a straight line relationship holds, this is exactly true for every choice of a. If the true relationship is nonlinear, then β_1 represents some sort of average difference.

2. Similarly, we know that every c unit difference in X is associated with a $c\beta_1$ unit difference in $\log(\theta_x)$:

$$\log(\theta_{a+c}) - \log(\theta_a) = (\beta_0 + \beta_1 \times (a+c)) - (\beta_0 + \beta_1 \times a) = c\beta_1.$$

3. We do not find it very convenient to talk about $\log(\theta)$, however. We would rather talk about θ (i.e, the odds, the mean, or the hazard). Hence we back transform to obtain statements about the *ratio* of θ across groups. So we find that every 1 unit difference in X is associated with a e^{β_1}-fold change in θ:

$$\frac{\theta_{a+1}}{\theta_a} = e^{\beta_1} \qquad\qquad \frac{\theta_{a+c}}{\theta_a} = e^{c\beta_1} = \left(e^{\beta_1}\right)^c.$$

   We can similarly say that the odds ratio (in logistic regression), the mean ratio (in Poisson regression), or the hazard ratio (in proportional hazards regression) is e^{β_1} for each 1 unit difference in the value of X, and the ratio is $e^{c\beta_1}$ for each c unit difference in the value of X.

7.2 Transformations of Predictors

Suppose we model

$$\log(\theta_x) = \beta_0 + \beta_1 \times \log_k(x)$$

1. From our standard interpretation of regression slope parameters, we know that every 1 unit difference in $\log_k(X)$ is associated with a e^{β_1}-fold change in the value of θ, and every c unit difference in $\log_k(X)$ is associated with a $e^{c\beta_1}$-fold change in the value of θ.

2. In units of X, we know that a 1 unit difference in $\log_k(X)$ is a k-fold increase in X, and similarly a c unit difference in $\log_k(X)$ is a k^c-fold increase in X . If k is some convenient multiple (i.e., a doubling when $k = 2$) we just say that for every k-fold increase in X the value of θ increases e^{β_1} fold.

3. Note that in Stata and R, the output for logistic, Poisson, or proportional hazards regression can be provided on either the scale of the β's or as e^{β}. You have to keep track of which is which.

- In Stata, `logit` returns the β's (and includes the intercept β_0), and `logistic` returns the e^{β}'s (and suppresses the intercept).
- In Stata, `poisson` returns the β's (and includes the intercept β_0) by default. If you specify option `irr`, Stata returns the e^{β}'s (and suppresses the intercept).
- In Stata, `stcox` returns the e^{β}'s by default, and if you specify option `nohr` Stata returns the β's. (Note that the baseline hazard function takes on the role of an intercept in proportional hazards regression, and is never returned as part of the standard regression output.)
- In R, summaries of the output of glm() and coxph() will tend to give both.

8 Communicating Ratios in Natural Language

It is often quite difficult for people to interpret the many ways that we might talk about ratios. Below I present several examples of how you might describe the output when using a log link with an untransformed predictor. I will use the example of looking at the odds of "response" as a function of dose measured in grams.

1. For an estimated odds ratio of 1.31 I might say any of
   (a) "the odds of response is 1.31-fold higher in the experimental group taking 1 g of drug than it is in the control group taking placebo" *(I tend to use this phrasing when there are only two groups. In this case I would also tend to explicitly state the odds (and/or probability) of response in each group, unless there were other covariates in the model.)*
   (b) "the odds of response is 1.31-fold higher for every 1 g difference in dose"
   (c) "the odds of response is 31% higher for every 1 g difference in dose" *(I tend to prefer this one for $1 < OR < 2$)*
2. For an estimated odds ratio of 2.31 I might say any of
   (a) "the odds of response is 2.31-fold higher in the experimental group taking 1 g of drug than it is in the control group taking placebo" it (I tend to use this phrasing when there are only two groups. In this case I would also tend to explicitly state the odds (and/or probability) of response in each group, unless there were other covariates in the model.)
   (b) "the odds of response is 2.31-fold higher for every 1 g difference in dose" *(I tend to prefer this one for $OR > 2$)*
   (c) "the odds of response is 131% higher for every 1 g difference in dose"
3. For an estimated odds ratio of 0.91 I might say any of
   (a) "the odds of response in the experimental group taking 1g of drug is 9% lower than the odds in the control group taking placebo" *(I tend to use this phrasing when there are only two groups. In this case I would also tend to explicitly state the odds (and/or probability) of response in each group, unless there were other covariates in the model.)*
   (b) "the odds of response is only 0.91 times as high for every 1 g difference in dose" *(I tend to prefer this one for $OR < 1$ when there are more than two groups)*
   (c) "the odds of response is 0.91 times as high for every 1 g difference in dose"
   (d) "the odds of response is 9% lower for every 1 g difference in dose" *(I tend to think this is a little more confusing, because when you are going to consider a difference in dose of c grams, you have to take 0.91^c, rather than use 0.09)*

Note the asymmetry of ratios: If the experimental to control odds ratio is 1.25 (so 25% higher), the control to experimental odds ratio is 0.80 (so 20% lower).