Biost 517, Emerson
Supplement #2
January 12, 2005

Use of Logarithmic Transformations

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. 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. $\log_k(x^y) = y * \log_k(x)$

Logarithmic Transformations in One and Two Sample Problems

Suppose we have random variables X_i and Y_i. If we take logarithmic transformations $W_i = \log_e(X_i)$ and $Z_i = \log({}_eY_i)$, 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.

Logarithmic Transformations in Regression Models

Transformations of Predictors

Suppose we model

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

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

Transformation of Response

Suppose we model (for arbitrary base j)

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

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

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

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 assumptions of linear regression, 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

$$mdn[\log_j(Y)] = \log_j(mdn[Y]) = \beta_0 + \beta_1 \times X$$

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.

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)

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

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.