Logarithms and Their Use in Regression

Kurt M. Anstreicher

Logarithms have mathematical properties that turn out to be very useful in the construction of simple and multiple regression models. This note summarizes important properties of logarithms and shows how they are applied in the regression context.

1 Definitions

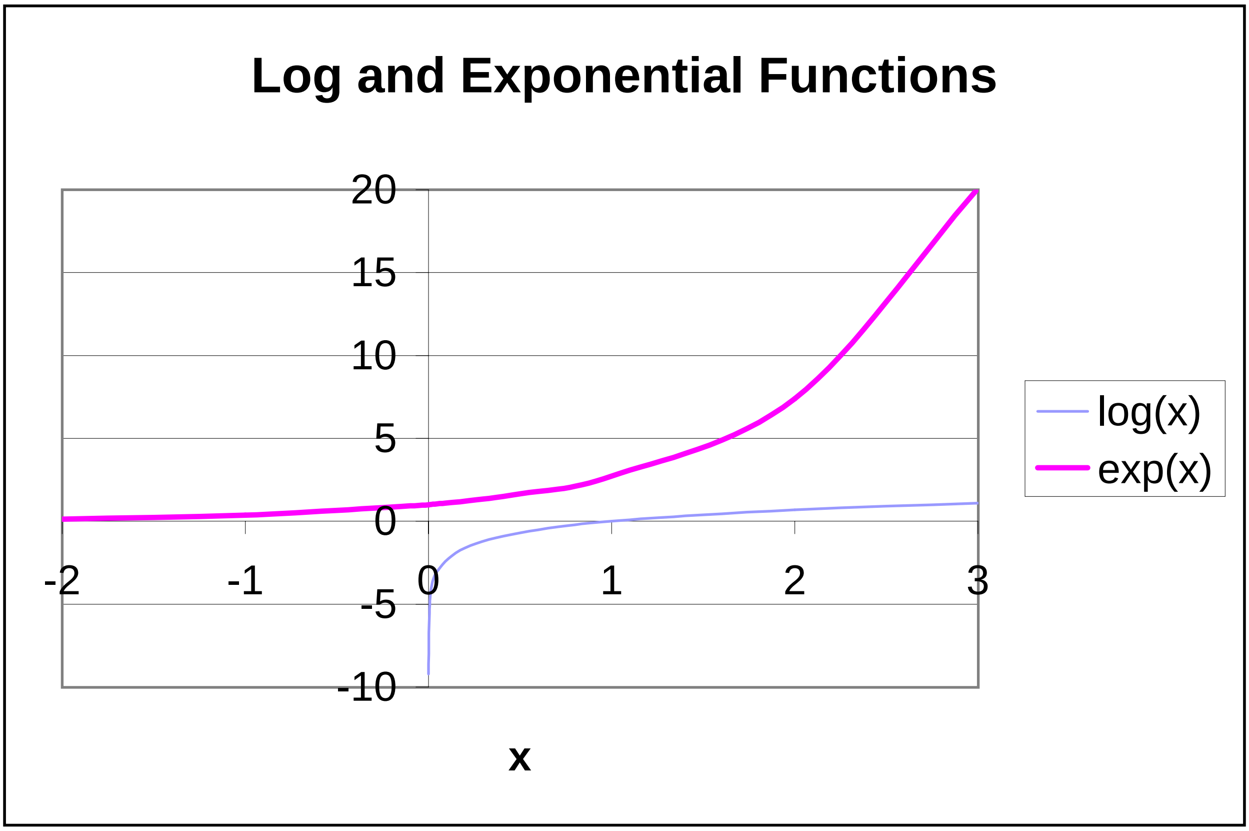
The logarithm is defined via the relationship

$$\log_a(x) = p \iff a^p = x,$$

where $a > 1$, $x > 0$ and the equation on the left reads "log to the base a of x equals p." The most common choices for the base are 2, 10, and e, a special number equal to about 2.71828. Some examples are:

$$\begin{aligned}
10^2 = 100, &\quad \text{so} \quad \log_{10}(100) = 2;\\
10^{-1} = .10, &\quad \text{so} \quad \log_{10}(.10) = -1;\\
10^{2.5} \approx 316.22766, &\quad \text{so} \quad \log_{10}(316.22766) \approx 2.5;\\
\\
2^5 = 32, &\quad \text{so} \quad \log_2(32) = 5;\\
2^{-2} = .25, &\quad \text{so} \quad \log_2(.25) = -2;\\
2^{7.4} \approx 168.8970, &\quad \text{so} \quad \log_2(168.8970) \approx 7.4;\\
\\
e^{2.302585} \approx 10, &\quad \text{so} \quad \log_e(10) \approx 2.302585\ .
\end{aligned}$$

The logarithm to the base e, or $\log_e(\cdot)$, is often refered to as the "natural log," and denoted $\ln(\cdot)$. The function e^x is also often written $\exp(x)$, and is called the "exponential function." In regression applications the natural log is used almost exclusively, so henceforth $\log(\cdot)$ means $\log_e(\cdot)$. The (natural) logarithm and exponential functions are illustrated in a figure below.



2 Doubling time and half life

The exponential function e^{bt} with $b > 0$ represents *exponential growth,* and with $b < 0$ represents *exponential decay.* A typical example of exponential growth is the investment of some amount D which then generates interest at an annual rate b, componded continuously. In this case the value of the investment at time t (in years) is De^{bt}. For example, if $b = .05$ then the interest rate is 5%, and the value of the investment after one year would be $e^{.05}D \approx 1.0513D$, corresponding to an "annual yield" of about 5.13%. A natural question to ask is how long it then takes for the value of the investment to double. This requires $e^{bt} = 2$, which is the same as saying that $\log_e(2) = bt$, or $t \approx .693/b$. Therefore the doubling time is about equal to 70 divided by the interest rate, in percentage points. For example if the interest rate is 7%, compounded continuously, then the doubling time is about 10 years.

A common example of exponential decay is the deterioration of radioactive compounds, or pollutants. In this case there is an initial quantity Q, and after time t the amount that remains is Qe^{bt}, where $b < 0$ and $|b|$ is called the decay rate. A natural question in this case is how long it takes for the quantity present to be reduced by one-half. This is the same as requiring $e^{bt} = 1/2$, which is $\log_e(1/2) = bt$, or $t \approx -.693/b = .693/|b|$. Therefore the half-life is equal to about 70 divided by the decay rate, in percentage points. For example if the decay rate is 10%, then the half-life is about 7 years. This also means that after 21 years the quantity present would be about 1/8 the original quantity.

3 Properties

There are a small number of important properties of the log and exponential functions, summarized below. These properties, or "rules," are used very extensively when evaluating expressions involving logs and exponentials.

1. *Inverse Rules.* $\log(e^x) = x$, and $e^{\log(x)} = x$.
2. *Product Rules.* $e^x e^y = e^{x+y}$, and $\log(xy) = \log(x) + \log(y)$.
3. *Power Rules.* $(e^x)^y = e^{xy}$, and $\log(x^y) = y \log(x)$.

Note that since $e^0 = 1$, the first rule implies that $\log(1) = 0$. Another useful property of e^x and $\log(x)$ relates to how these functions behave near 0 and 1, respectively. In particular,

$$e^x \approx 1 + x, \qquad \log(1 + x) \approx x,$$

and these approximations are reasonably good for $-.2 \leq x \leq .2$. For example $e^{.05} \approx 1.0513$, $e^{.10} \approx 1.1052$, $e^{.20} \approx 1.2214$.

4 Use of logs in regression

By using logs some common nonlinear relationships can be "transformed" into linear equations. This is very convenient in certain regression applications.

4.1 Log-linear regression

Consider an exponential relationship between two variables X and Y,

$$Y = Ae^{bX},$$

where $A > 0$. Taking the log of both sides, using the rules described above, results in the equivalent relationship

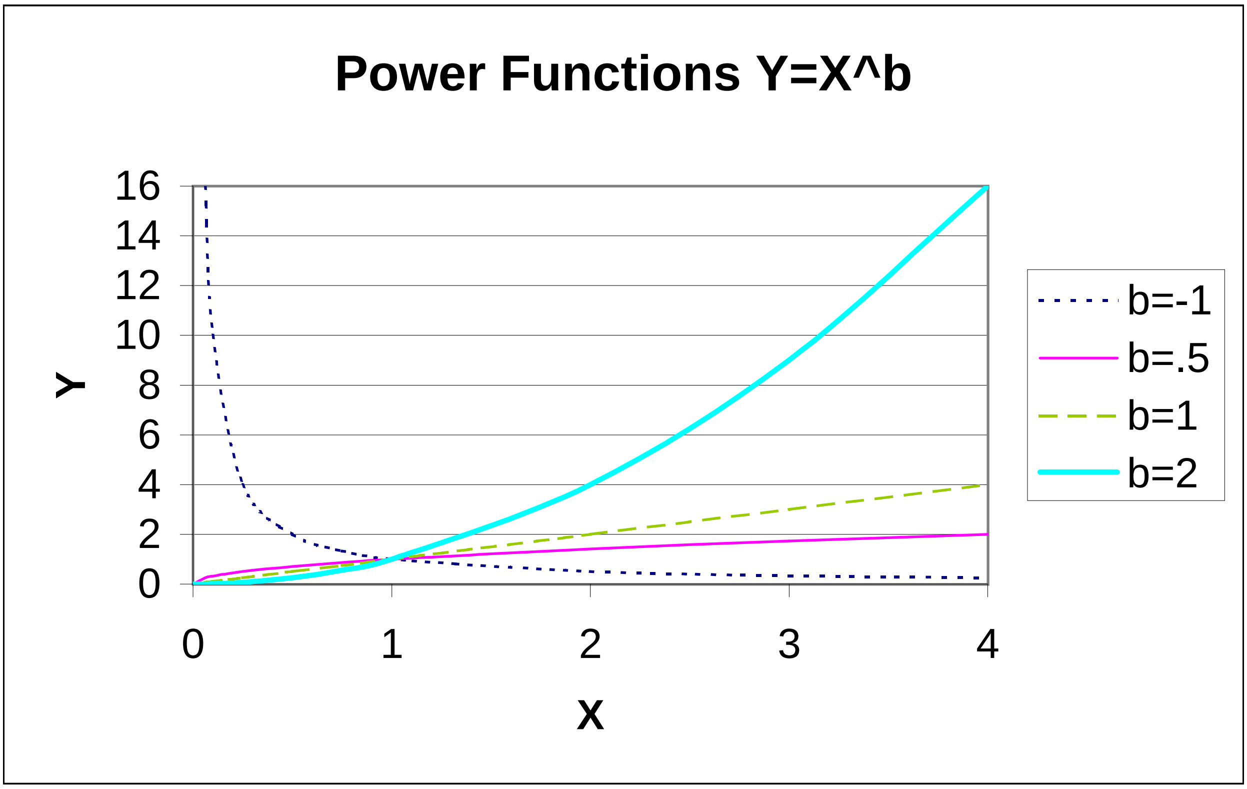
$$\log(Y) = a + bX,$$

where $a = \log(A)$. Note that this has the form of a simple regression equation, relating the variables $\log(Y)$ and X. This type of relationship is often called *log-linear*, or *semi-log.*

Example. Consider Y to be the Consumer Price Index (CPI) of the United States from the years 1946 to 1999, based on an index value of 100 from 1983. If t is taken to be the year, coded $1, 2, \ldots, 54$, a regression of log(CPI) on t produces the equation

$$\log(\text{CPI}) = 2.8001 + .04357t.$$

The intercept and slope in this equation have simple interpretations. The intercept is the value of log(CPI) at $t = 0$, corresponding to 1945. In other words the CPI for 1945 would be about $e^{2.8001} \approx 16.45$. The slope coefficient is the growth rate, with continuous compounding.



This means that the average annual inflation rate in the period 1946-1999 is approximately 4.36%. More exactly, exponentiating both sides of the regression equation results in the equation

$$\text{CPI} = e^{2.8001}e^{.04357t} \approx 16.45(e^{.04357})^t \approx 16.45(1.0445)^t,$$

so the average inflation rate is 4.45%.

In general, in a log-linear regression equation $\log(Y) = a + bX$, the slope coefficient b gives the approximate relative change in Y per unit increase in X, and this approximation is reasonably good for relative changes up to $\pm.2$, or 20%.

4.2 Log-log regression

Next consider a *power* relationship between X and Y of the form

$$Y = AX^b,$$

where $A > 0$. Some examples (with $A = 1$) are illustrated in the figure above. Taking the logs of both sides, using the rules above, results in the equivalent equation

$$\log(Y) = a + b\log(X),$$

where $a = \log(A)$. The transformed equation has the form of a simple regression equation relating the variables $\log(X)$ and $\log(Y)$. This is commonly refered to as a *log-log* relationship.

Example. Let X be the total number of votes cast, and Y the number of votes cast for Pat Buchanan, in Florida counties in the 2000 US presidential election, with the data for Palm Beach county omitted. A regression of $\log(Y)$ on $\log(X)$ results in the regression equation

$$\log(\text{Buchanan}) = -2.50207 + .70349 \log(\text{Total}).$$

Exponentiating this equation, using the fact that $e^{-2.50207} \approx .082$, results in the relationship

$$\text{Buchanan} = .082\text{Total}^{.70349} .$$

Note that as the total number of votes cast goes up by a factor of 10, the number of votes for Buchanan goes up by the factor $10^{.70349} \approx 5$, so that Buchanan's share is cut in half.

In general, in a log-log regression equation $\log(Y) = a + b\log(X)$, the slope coefficient b can be interpreted as giving the approximate percentage change in Y per percentage change in X. (In economics this is commonly refered to as an *elasticity.*) This approximation is reasonably good if the percentage changes in both X and Y are no more that ±20%.

4.3 Standard error of estimate

In simple regression the standard error of estimate, denoted S_e, estimates the variation in Y about its mean (equal to its median) at any given X. When the dependent variable is $\log(Y)$, S_e is still a measure of the variation in Y, but this variation is *proportional* to the median value of Y at a given X. To see this, consider the "model" equation in the log-linear case (the log-log case is very similar)

$$\log(Y) = \alpha + \beta X + \epsilon,$$

where ϵ is assumed to be a normal random variable, with mean 0, that is independent of X. Exponentiating both sides results in the equivalent equation

$$Y = e^{\alpha+\beta X} e^{\epsilon}.$$

Since zero is the median value of ϵ, and $e^0 = 1$, the median value of Y at a given X is $e^{\alpha+\beta X}$. Nonzero values of ϵ result in deviations in Y from this median value. For example, from the assumed normality of ϵ we know that

$$P(e^{\alpha+\beta X} e^{-\sigma_\epsilon} \leq Y \leq e^{\alpha+\beta X} e^{\sigma_\epsilon}) \approx .68.$$

Moreover if σ_ϵ is not too large, then $e^{-\sigma_\epsilon} \approx 1 - \sigma_\epsilon$, and $e^{\sigma_\epsilon} \approx 1 + \sigma_\epsilon$. For example, for $\sigma_\epsilon = .1$ we have $e^{-.1} = .905$ and $e^{.1} = 1.105$, resulting in

$$P(.905 e^{\alpha+\beta X} \leq Y \leq 1.105 e^{\alpha+\beta X}) \approx .68 .$$

In other words, 68% of the time Y is within about 10% of its median value, for any given value of X. Since S_e is an estimate of σ_ϵ one can similarly think of S_e as measuring the *relative* variation in Y, about its median, at any given X. In addition, because S_e usually dominates forecast error, an approximation of a 95% forecast interval for Y at any given X is given by the interval $[e^{a+bX} e^{-2S_e}, e^{a+bX} e^{2S_e}]$. Therefore $2S_e$ is the approximate relative error associated with a 95% forecast interval.

5 Multiple regression

Consider a multiple regression equation of the form

$$\log(Y) = a + b_1 \left\langle \begin{array}{c} X_1 \\ \log(X_1) \end{array} \right\rangle + b_2 \left\langle \begin{array}{c} X_2 \\ \log(X_2) \end{array} \right\rangle + \ldots + b_k \left\langle \begin{array}{c} X_k \\ \log(X_k) \end{array} \right\rangle ,$$

where the notation

$$\left\langle \begin{array}{c} X_i \\ \log(X_i) \end{array} \right\rangle$$

indicates that either the variable X_i or $\log(X_i)$ is used. As in the case of ordinary multiple regression, to interpret an individual slope coefficient b_i one thinks of changing X_i while holding all other X_j, $j \neq i$ fixed. However, for any set of fixed values for the other variables, the relationship between Y and X_i then takes the form

$$\log(Y) = \hat{a} + b_i \left\langle \begin{array}{c} X_i \\ \log(X_i) \end{array} \right\rangle ,$$

where $\hat{a}$ is determined by a and the values of the X_j, $j \neq i$. In other words, to interpret the effect of an individual variable X_i, for given values of the other variables, one can think of a *simple* log-linear or log-log equation that relates Y and X_i. The interpretation of S_e in the case of a multiple regression with dependent variable $\log(Y)$ is also exactly the same as described in the case of simple regression above; the only difference is that the median value of Y is determined by the values of all the independent variables $X_1, X_2, \ldots, X_k$.

Example. The "Artsy" data set consists of observations of the following variables for 256 employees of a company.

SEX: Sex of employee, coded 0 for female, 1 for male.
GRADE: Grade of employee's job in 1981; values are 1–8.
TINGRADE: Number of years employee has been in current grade.
RATE: Weekly pay rate as of 12/31/81.

A regression restricted to employees in grades 6, 7, and 8 results in the equation

$$\log(\text{RATE}) = 5.7938 + .0996\,\text{SEX} + .0153\,\text{TINGRADE} + .0546\,\text{GRADE7} + .1901\,\text{GRADE8},$$

where GRADE7 and GRADE8 are dummy variables indicating that an employee is in grade 7 or grade 8, respectively. (The "base case" for the regression therefore corresponds to an employee in grade 6). The intercept and slope coefficients in this regression have simple and very useful interpretations. The exponentiated value of the intercept, $A = e^a = e^{5.7938} \approx 328.26$, is the average weekly pay rate for a woman who has just started working in grade 6. The coefficient for TINGRADE, .0153, implies that the average annual pay raise is approximately 1.5%. The coefficients for GRADE7 and GRADE8 give the approximate relative increases in starting pay in these grades compared to the starting wage in grade 6. For example, the average starting salary in grade 8 is about 19% higher than in grade 6. (A more accurate figure is obtained by computing $e^{.1901} = 1.21$, implying that the average starting salary in grade 8 is about 21% higher than in grade 6.) The coefficient of the SEX variable gives the typical relative increase in salary when comparing a male employee to a female, with equal values of the other variables. The regression therefore estimates that in grades 6, 7, and 8 men are paid approximately 10% more than women, for equal values of GRADE and TINGRADE. The standard error S_e in this regression is .0931. Using the assumption of normally distributed residuals, this implies that the error in the fitted versus observed value of RATE is no more than ±9.3% for approximately 68% of the observations, and no more than ±18.6% for approximately 95% of the observations.