

Log Lines (Changes, of Logs)

Logs: Once again, as with averages and lines, you know the math. of what we are about to discuss. *Using* the math is another matter and that is my subject.

What is a logarithm? There are several ways of defining it that come to the same thing. Let's take the secondary school approach (as compared to the calculus/college approach) because it is probably the useful one for data analysis. The story of logarithms begins with the story of exponents. We know for example that in some cases the multiplication of two numbers in exponential form can be seen as addition, addition applied to the exponents. Thus, if one number is 2^3 , two cubed, and another is 2^4 , two to the fourth power, then the multiplication of these two numbers can be expressed, equally well as multiplication, multiplication of the numbers themselves, or as addition, addition of their exponents. That is = 8

	Exponential Form	=	Standard Form
First number:	2^2	=	4
Second number:	2^3	=	8
Third number:	2^5	=	32

Product, written as a multiplication problem:

$$4 * 8 = 32$$

Product, written as addition problem (adding exponents):

$$(2^2)*(2^3) = 2^{2+3} = 2^5 = 32$$

Or, in symbols

$$b^p * b^q = b^{p+q}$$

That is a quick review of “high school” mathematics of exponents, with perhaps one more useful fact that should be remembered: What you have, above, is that multiplication, applied to the original numbers, can (as above) be re-expressed as addition, applied to the exponents. This basic correspondence between multiplication of simple numbers and addition of exponents implies another correspondence between exponentiation, of the simple numbers and multiplication of the exponents

Operation applied to
the simple numbers

Operation applied to
the exponents

Multiplication ----->

Addition $b^p * b^q$ -----> b^{p+q}

Exponentiation ----->

Multiplication $(b^p)^r$ -----> (b^{rp})

These properties of exponents depend on an implicit assumption, specifically, that the “thing” being exponentiated stays the same from number to number: Above, when I worked with numbers, the “thing” was always 2, I always used exponents of 2. Here, symbolically, the thing is b, I have used exponents of “b”. It doesn’t matter what number I exponentiate as long as I am consistent. The number I use is called the base. and its exponent is the logarithm of x base b.

$$b^{\log_b X} = X$$

You are accustomed to seeing these things in two forms: In logs base 10, called “common logs”, and particularly in the sciences and mathematics in logs base “e”, where it is useful to use a special constant e, approximately equal to 2.71828. A third relatively rarely-used base is simply 2 (often used in information theory and related areas of computer science). Thus

Number	Log Base 10 “Common Logs”	Log Base e “Natural Logs”	Log Base 2
1	.0	.0	.0
1.01	.0043	.01	.014355
1.02	.0086	.02	.028569
1.03	.0128	.03	.042644
2	.3010	.69	1.
4	.6021	1.39	2.
8	.9031	2.08	3.
10	1.	2.30	3.321928
100	2.	4.61	6.643856
1000	3.	6.91	9.965784

Before we use these things in data analysis, let me dwell a little on the actual values. Why, for example, would you prefer one base to another? Mathematically, it makes no difference. Mathematically, these are just three different ways of doing the same thing — three ways of taking advantage of the properties of exponents. But if not mathematically, then practically it makes a difference for the usual reason: Appropriate numbers serve to direct attention to regularities in the data. Note, for example, that each of these bases for the logarithm gives you easily remembered numbers in some range, less easily remembered numbers in other ranges. So, if you are using data in which binary arithmetic is important (as for computers), use

logs base 2: It takes $\log_2(x)$ bits (rounded up), to record the value of an integer.

If you are using data in which items vary by orders of magnitude (i.e., by powers of 10), then use logs base 10: In 1990 the population of the United States was approximately 250 million (250,000,000), or “two times ten to the eighth”. (Counting digits to the left of the decimal point minus one.) This is sometimes referred to as “scientific notation”, “ 2.5×10^8 ” — where you recognize “8” as the logarithm, base 10 (approximately). The population of Canada was “two times ten to the seventh”. Ah, I know immediately that the United States population is an order of magnitude larger than the population of Canada. Base 10 helps. Using the raw number, “26,538,000”, it is mentally clumsy to compare it to other numbers. Remember it as approximately 10^7 and you know immediately that it was an order of magnitude smaller than the size of the United States.

Base e, is very convenient when you are working with small percent increases, as in economics and in population data for which an economy, or a bank account, or a population is likely to grow at a rate of between 1 and 10 percent per year. It is convenient because of a transparent correspondence between ratios and their logarithms when the ratios are in this range. In this range I have need for log tables, I can do it “in my head”. For ratios close to 1

Ratio	Ln of the Ratio (Logarithm, base e) (Two digits)	(Four Digits)
1	0	0

1.01	0.01	0.0100
1.02	0.02	0.0198
1.03	0.03	0.0296
1.04	0.04	0.0392
1.05	0.05	0.0488
1.06	0.06	0.0583
1.07	0.07	0.0677
1.08	0.08	0.0770
1.09	0.09	0.0862
1.10	0.10	0.0953
1.11	0.10	0.1044
1.12	0.11	0.1133
1.13	0.12	0.1222
1.14	0.13	0.1310

Within this range, computing logarithms, base e, is easy. In turn this makes it simple to work with compound interest on interest rates within this range. For example, let me do some compound interest “in my head” asking how long it takes for the principle to double at various rates of interest. The rough answer is what accountants call “The rule of 70.” Which means take the interest rate, divide 70 by the interesting rate. That is the doubling time. Or, you can ask the question, if I know the doubling time, then what is the interest. The logic is straightforward:

Suppose I start with a principal of \$100. At one percent growth per year, how long will it take to double? At the end of one year I will have \$100*1.01. At the end of two years I will have \$100*1.01*1.01. How many years in a row do I have to apply the multiplier 1.01 before I get \$200?

$$\$100 * 1.01 * 1.01 * 1.01 * \dots * 1.01 = \$200$$

If I use logs base e, and if commit to memory that the log of 2 is approximately .70, then the “trick” is to convert this multiplication problem to an additon problem:

In logs, How many years in a row do I have to add .01 before it adds up to ln 2.

$$\ln(100) + \ln(1.01) + \ln(1.01) + \ln(1.01) \dots\dots\dots + \ln(1.01) = \ln(100) + \ln(2)$$

combining terms

$$\ln(100) + n(\ln(1.01)) = \ln(100) + \ln(2)$$

This is the logarithmic equivalent of asking how many multiplications by 1.01 (how many additions of ln 1.01) do I need to achieve doubling. How do I solve it? I remove the troublesome log(100), no need for it

$$n(\ln(1.01)) = \ln(2)$$

I pluck the logarithms from my memory. Substiuting the values of the logarithms into the equation:

$$n(0.01) = .70$$

And now I know:

$$n = 70 \text{ (approximately)}$$

At 1% per year, the principal will take approximately 70 years to double. How long will it take for my money to double at 2% per annum? In detail:

$$n(\ln(1.02)) = \ln(2)$$

Inserting the logarithms, base 3:

$$n(0.02) = .70$$

And therefore, approximately,

$$n = 35 \text{ (approximately)}$$

At 2% per year, the principal will double in approximately 35 years.

How many years will it take for my money to double at 10% per annum? In detail:

$$n(\ln(1.10)) = \ln(2)$$

Inserting the logarithms, from memory:

$$n(0.10) = .70$$

And now I know

$$n = 7 \text{ (approximately)}$$

At 10% per year, the principal will double in approximately 7 years.

When would I use such things, other than to impress my bank by doing compound interest in my head? Frequently. For example, we are about to analyze the rate of growth of the population of the United States. In two hundred years, from 1790 to 1990, it grew from approximately 3 million people to approximately 250 million people. So: What was the average annual rate of increase? As usual, I can do a good approximation to the full data analysis quickly, and in my head:

Here's the logic: I focus on doubling. How many times has the U.S. population doubled? That's, 3 million, 6 million, 12 million, 24 million, 48 million, 96 million, 192 million, 384 million: It doubled a little more than 6 times in 200 years. So, roughly, it doubled every 33 years. Ah: If it doubled in 33 years, then the average annual rate must have been approximately 2%.

I've barely begun to analyze these data, but I've got a baseline. The population increase was approximately 2% per year.

I put it to you:

In 1945 the U. S. federal government took in \$45 billion dollars in revenue. In 1992 the federal revenue was \$1.075 trillion dollars. What was the average annual rate of increase?

In 1945 the U.S. federal outlays for national defense were 82 billion dollars. In 1992 the outlay was 307 billion. What was the average annual rate of increase?

In 1945 the U.S. federal outlays for human resources were 2 billion dollars. In 1992 the outlay was 777 billion. What was the average annual rate of increase? (Source, Statistical Abstract of the United States, 1992, Table 491, page 315.)

In 1960 U. S. national health expenditures added up to 27.1 billion dollars. In 1990 the total was 666 billion. What was the average annual rate of increase?

In 1960 U. S. national health expenditures added up to 143 dollars per capita. In 1990 the figure was 2,566 dollars per capita. What was the average annual rate of increase? (Source, same. Table 135, page 97)

In 1959-60 U. S. personal income per capita was approximately \$2,200. In 1990 the figure was 18,720 dollars . What was the average annual rate of increase? (Source, same. Table 678, page 431)

The Slide Rule

For about 300 years most scientists had an intuitive grasp of logarithms as an indirect consequence of using something called the slide rule. So, as a curiosity and, to aid your intuition, let's discuss the slide rule.

I assume that the original reason for their use was the computational simplicity they introduced for multiplication — when you haven't yet invented the computer you care very much about such computational simplicity. I'm guessing, but I'd make a small bet that this device goes back to at least the fifteenth century, Regiomontanus???, and stayed in very heavy use until at least the 1970's. Today there is no need for them. But we do need the intuitive comfort with logarithms that the slide rule created. Assuming that they are not even manufactured any more, offer you as a paper cut out slide rule — serviceable but somewhat fragile.

The slide rule is a machine that adds logarithms physically by adding-up physical lengths. Starting with the obvious, start with 4 times 8. Multiplying 4 times 8 is simple. Multiplying 4 times 8 using logarithms, you add $\log 4$ to $\log 8$ and get an answer which is equal to $\log 32$.

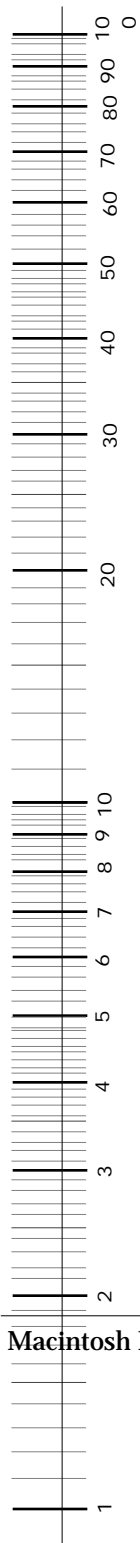
On the slide rule you do the same thing by adding the length corresponding to 4 to the length corresponding to 8 and reading that the result is the length corresponding to 32:

Find the "4" on the first scale.

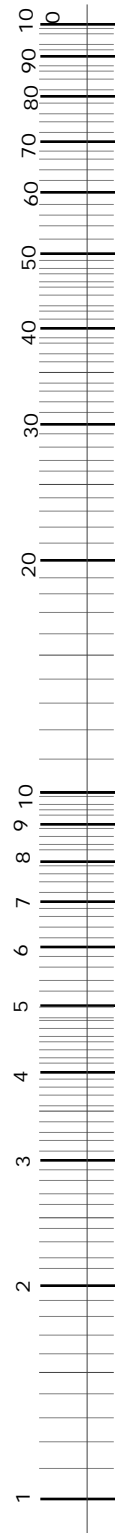
Place the "1" on the second scale next to the "4" on the first scale

Find the "8" on the second scale,

And then read the answer on the first scale (finding "32" opposite the "4")



Two-Scale Slide Rule:
Tear into two strips and lay them edge to edge



The slide rule creates intervals for 1, 2, 3 that are the correct intervals when numbers are related by multiplication. Check it: Use a ruler to measure the distance between 1 and 4 on the slide rule and then measure the distance between 1 and 8 on the slide rule. You will find that the physical distance from 1 to 8 is 50% larger than the physical distance from 1 to 4 just as, in the exponents above, the exponent for 8 (as in $2^3 = 8$) is 50 percent larger than the exponent for 4 (as in $2^2 = 4$).

Linear Relations Using Logs

Interpreting $\text{Log } y = mx+b$

And now, finally, what happens if well-behaved form of a variable is its logarithm and you actually have to logarithms with data? You've got a variable: It's asymmetrical in its original form. But when you transform it, using logs, it is symmetrical. You look at the variable: It's dollars (an amount in Tukey's language) — you expect it to need logs. You plot the variable with respect to another variable and the result is definitely not linear — but when you plot its logarithm with respect to the other variable the result is definitely (close to) linear. So, the message is — “use logs”, but how do you interpret such things?

To discuss the “meaning” of something whose logarithm appears to be a linear function of another variable — called a semi-log (or semi-logarithmic) relation — there is no trick: You have all of the mathematical tools, it's just a matter of using them.

So, here's what you've got, a semi-logarithmic linear relation

$$\ln Y = m X + b$$

I know from simple math that if the semi-log relation is true, in nature, then an exponential relation is also true in nature. The second equation helps to interpret the first:

If $\ln y = mx + b$

then $y = e^{mx+b}$

Intercept

First, what is b? In the semi log equation, b is the intercept. Taking the algebra a step further, the form of the equation in dollars is

$$y = B e^{mx}, \text{ where } B = e^b$$

This demonstrates that the addition of b in the semilog relation implies a proportionality to anti-log b in the second equation. It says that y is proportional to an exponential function of x, where the proportionality, upper-case B, is the anti-log of lower-case b in the linear equation.

Slope

Now, second, what is m? In the linear form, of course, it means that $\ln y$ increase up by m units for each unit increase in x — that's what a slope says in a linear relation. If that is what happens in the logarithmic equation, what happens in the second equation for plain dollars — without logs? There is no secret to answering the question. You just add 1 to x and see what happens.

So, in simple terms I ask again, what happens as x goes up by one unit? I add one and look at the results

$$y = Be^{mx}$$

$$y = Be^{m(x+1)}$$

That express the new value of y as y' , corresponding to an x that is increased by 1. Simplifying the second equation, I get

$$y' = B e^{mx+m}$$

Breaking the exponential factor into two factors instead of just one, that is equivalent to

$$y' = B e^{mx} e^m$$

And now I recognize that the first part of the stuff on the right matches the original equation. So comparing y' to y the result depends on m . That tells me how to interpret m .

$$\frac{y'}{y} = \frac{Be^{m(x+1)}}{Be^{mx}}$$

$$\frac{y'}{y} = e^m$$

Ah, the effect of a unit increase in x is to multiply y by the value of e^m .

So suppose that m is a number like .03? I just pull out my calculator and figure out the value of the anti-log of .03. Or, if I am using logs base e , then I can remember the logs without computing them, remembering that $\exp(m)$ is equal to 1.03. Which means — here's the payoff, that y gets *multiplied* by 1.03 everytime x adds 1 to its value.

And finally, we wipe out the traces of what we've done and the way we actually analyzed the data, by using percentages —

because people feel comfortable with percentages — even though they aren't much use when you're doing the real work — and I announce: "Y increases at 3% per annum.)

Scatter

Finally, because this is data analysis, there is the third property of a data analyst's line: The scatter.

The residuals are departures from the semi-log equation, $\log(y) = mx + b$.

The trick is to put the residuals into a form that uses the words "plus or minus". If the residuals represent error, and nothing more, then the distribution of the residuals will be symmetrical around a mean of zero. So, for example, we can compute the standard deviation of the residuals and say, that the residuals have an average of zero "plus or minus" two standard deviations. Or, we can make the corresponding statement with medians and quartiles, saying that the residuals have an average of zero, "plus or minus" the number corresponding to the distance between the median and the quartiles (*either* quartile, since the distribution is symmetrical).

Then the corresponding statement in y (rather than $\log y$) is straightforward (except that the standard deviation in units of $\log y$ is not the same as the standard deviation in units of y — so we avoid the word standard deviation. So: The residuals show deviation within a factor of ___, where you fill in the blank with the antilog of two standard deviations of y. Or, using quartiles: Fifty percent of the predicted values lie within a factor of ___ above the predicted values of y and a factor of ___ below the predicted values, where you fill in the first blank with the anti-

log of the high quartile of the residuals and you fill in the second blank with the anti-log of the low quartile of the residuals. Or, using the inner fences: With few exceptions the predictions lie within a factor of __ above the predicted values and a factor of __ below them, where the blanks are filled in with anti-logs of the corresponding values of the fences of the residuals for log y.