The Neglected Logarithmic Graph

Presented to the Government Research Association Conference
July 25-27, 2011

Woods Bowman
DePaul University

This paper deals with two problems that confront analysts who must present data to the general public in ways that are technically precise, yet understandable to laypersons.

> Problem 1. Calculating the compound average growth rate (CAGR) of a variable over a period of time while taking account of events that may cause it to change abruptly from time-to-time, including identifying exactly when changes occur (i.e., the *break points*).
>
> Problem 2. Displaying data that differ by two or more factors of 10 (*orders of magnitude*) on a common chart.

Solutions in both cases make use of the convenient properties of logarithms (*logs*), which are easier to use and explain than is commonly recognized.

PROBLEM 1

Numerous economic variables, like prices and government spending, grow at a relatively constant rate over long periods and compound annually. The rate of change over time can be summarized by a single number: the CAGR. The formula for the CAGR depends on only two observations on a variable: its beginning value Y_t and ending value Y_{t+n}. It is easily described in a non-technical way by analogy to the interest rate on a savings account.[1]

$$CAGR = (Y_{t+n}/Y_t)^{(1/n)} - 1$$

Occasionally major exogenous events, such as recession, war, and natural disaster, disrupt the historic pattern. If the period of time being used in a trend analysis (the *framing period*) is sufficiently long, it is likely that any given variable, Y, will experience one or more disruptions in its CAGR and two data points may produce a misleading result. The CAGR will be too high to represent some sub-intervals and too low to represent others. Worse, it is not necessarily a "best" average value in the sense of minimizing the sum of squared deviations from the mean rate.[2]

To overcome this problem analysts break the framing period into sub-intervals where break points correspond to major events they *presume* to have caused it to change. They then calculate a CAGR for each sub-interval. However, this practice requires analysts to guess which events influence a significant change in growth rate and to guess the lag between cause and effect. This

---

[1] Despite its name, CAGR is not really an arithmetic average over the period n. Rather it is the rate that will result in Y_{t+n}, after "n" periods, starting from Y_t. The formula gives the growth rate in fractional terms. To convert to percent, multiply by 100%. Derivation of the formula is given in the appendix.

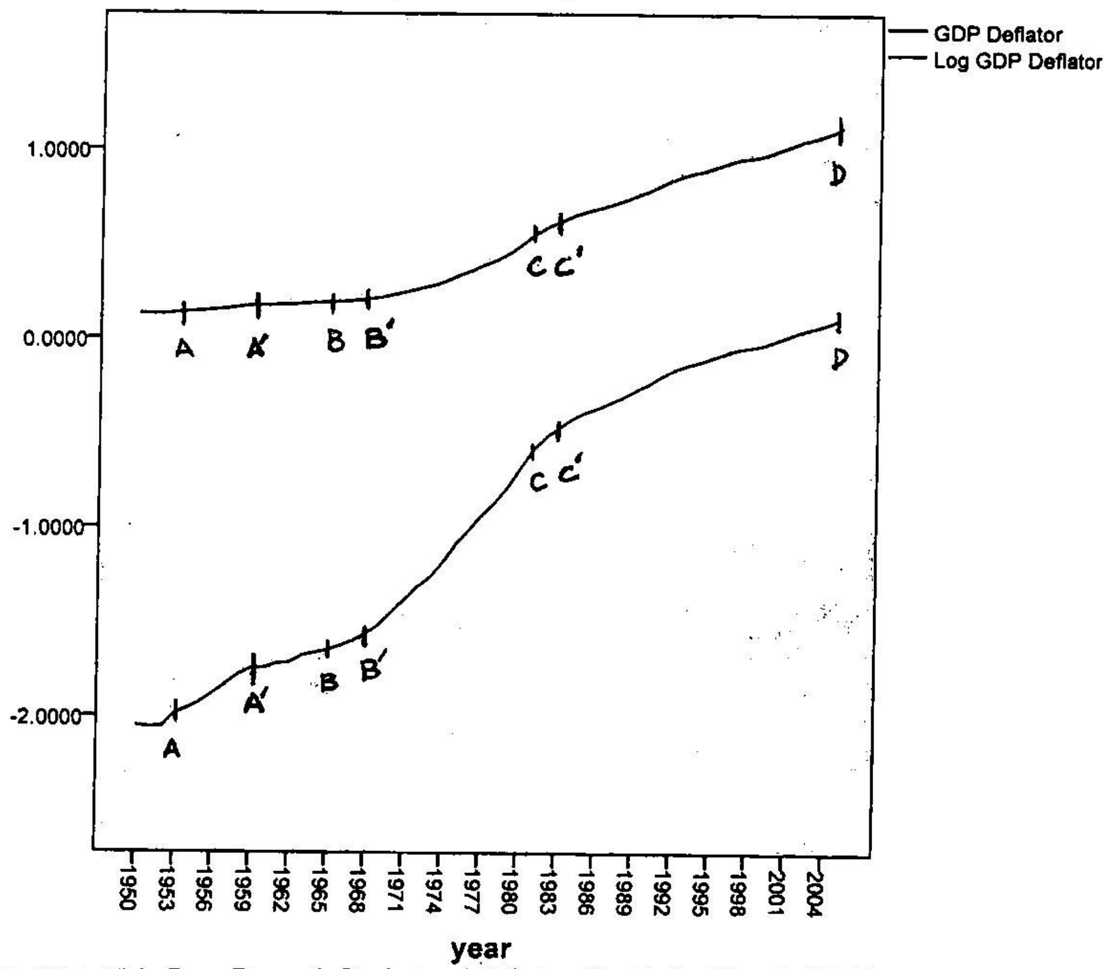
[2] This is the criterion for ordinary least squares (OLS) regression.

section describes a simple method that enables an analyst to find break points without any prior knowledge of disruptive events. The method requires plotting the log of a variable of interest, Y on a graph as in Chart A. Straight line segments correspond to sub-intervals where the CAGR is constant.[3] Page 8 of the appendix illustrates this proposition with computer-generated graphs of a smoothly compounding series and its logarithm.

Because intervals of constant CAGR are straight lines, actual break points are easier to identify. Analysts must still use visual identification of break points but logs eliminate optical illusions. As always, analysts should attempt to identify causal factors near the break points.

Chart A

GDP Deflator and Its Natural Log, 1950-2005
(dimensionless)



Source: Table 10.1: Gross Domestic Product and Deflators Used in the Historical Tables: 1940-2010., U.S. Office of Management and Budget (2000) *Budget of the United States Government, Fiscal Year 2000* http://www.gpoaccess.gov/usbudget/fy00/pdf/hist.pdf. Accessed April 6, 2011. (Note: Data between 2000 and 2010 are OMB estimates.)

[3] See appendix for proof and an example. I prefer the natural logarithm (base e) but any base will do.

The top line is the GDP price deflator. The points A, B and C correspond to historical markers. Different analysts might choose different markers but these points are likely candidates. (Ignore alternate points A′, B′, and C′ for the time being.)

- Point A ~ 1953 when an armistice ended combat in Korea. In this year President Eisenhower ended price and wage controls.
- Point B ~ 1965 when Congress passed his Great Society program, creating Medicare and Medicaid, *and* the Gulf of Tonkin resolution giving legal authority for a military buildup in Vietnam. Spending increases produced by this combination of war and domestic initiatives caused a sustained period of above-trend (runaway) inflation.
- Point C ~ 1981 when the prime rate reached 20%, finally damping inflationary pressures.

Using these event markers to calculate CAGRs gives the following results:

Table 1
CAGRs, 1953 to 2005

Points	Time Interval	N	CAGR
A-B	1953-1965	12	2.90%
B-C	1965-1981	16	6.80%
C-D	1981-2005	24	3.00%
A-D	1953-2005	52	4.10%

Source: see Chart A.

Break points in the vicinity of events A and B are far from obvious in the top line of Chart A. The bottom line is the natural log of the GDP price deflator. Comparison of the two lines reveals that runaway inflation did not begin until 1968 and it did not end until 1983.[4] Incorporating these changes in the set-up adds an additional break point as shown in Table 2. It is worth repeating that analysts must still use visual identification of break points but logs eliminate optical illusions

Table 2
CAGRs, 1953 to 2005, Alternate Set-Up

Points	Time Interval	N	CAGR
A-A′	1953-1960	7	3.60%
A′-B′	1960-1968	8	2.30%
B′-C′	1968-1983	15	7.50%
C′-D	1983-2005	22	2.70%
A-D	1953-2005	52	4.10%

[4] Indeed, inflation rarely starts or stops without lagging the proximate cause.

Source: see Chart A.

In Table 2, the period of runaway inflation is shifted forward slightly to B′-C′. The CAGR of the GDP deflator during this period is 0.7% points higher – which is reasonable.[5] The only material change in our conclusions comes in analysis of the period prior to 1968. Noting a break point at A′, we decompose this period into two sub-parts A-A′ and A′-B′ and calculate two CAGRs. Both are materially different from the 2.9% rate for the entire period leading up to runaway inflation in Table 1.

The question arises: what caused the reduction in CAGR in 1960? (To repeat: this graphical technique only helps to identify break points; it does not explain them.) A search of the historical record turned up nothing other than a change in presidential administration, so the most plausible explanation is that Korean War era wage and price controls had actually worked but excess demand for consumer goods accumulated. When controls were abandoned in 1953, excess demand caused prices to increase but by 1960 it had run its course and price increases moderated on their own.

PROBLEM 2

Large differences in scale between variables preclude presentation of their growth curves on the same chart using a common scale. However, when all data are expressed as logs, scale differences are rarely a problem. Chart B illustrates this point with logic transformations of data on total federal outlays and its defense and health components.

In 1956 real federal outlays were $70.6 bn, national defense outlays were $42.5 bn, but health outlays were a (comparatively) paltry $0.36 bn. By 2004, outlays for health, now including Medicare, had grown considerably. Total federal outlays were $2,479 bn, national defense outlays were $314 bn and health (including) Medicare outlays were $463 bn.[6]

The largest number ($2,479 bn) is nearly 7,000 times the smallest ($0.36 bn). Plotting the raw data in billions of dollars on the same scale would render early health outlays nearly invisible in the early years. Even having a divided scale with the total presented according to a left-hand scale and health plus Medicare presented according to a right-hand scale would not work well because the variation in the latter variable from 1956 to 2004 is nearly 1,300 times.

The only viable solution is to transform all data into logs and plot the results on a logic scale. This has the additional advantage, already discussed, that the slopes of the lines reflect their growth (or decay) rates. Chart B shows how this looks with the dimension of the vertical axis in billions of current dollars where ¾ of an inch equals a *multiple* of 12, 1.5 inches equals a multiple of 12^2, etc.

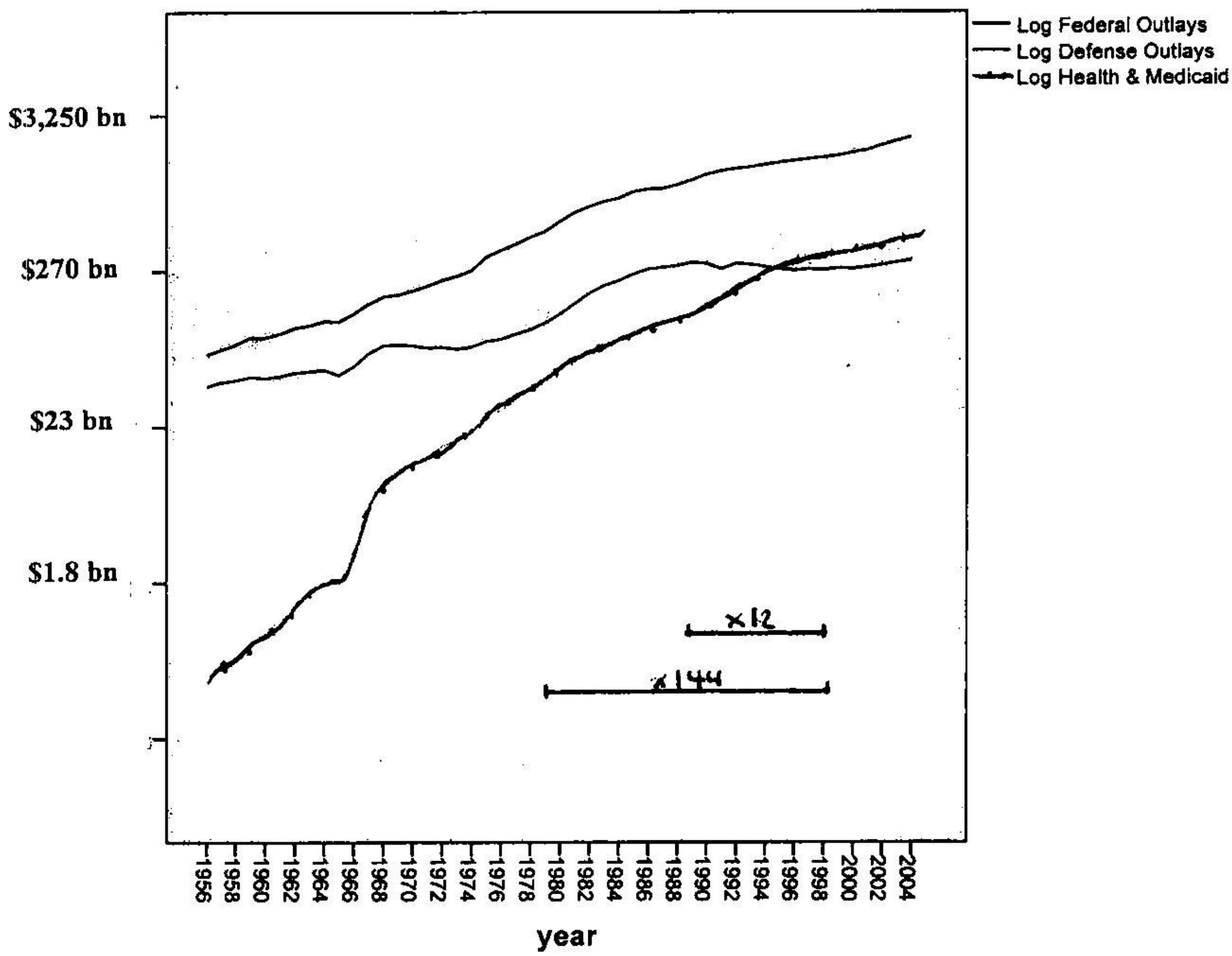
---

[5] The difference between CAGRs in Table 2 before and after 1983 is a full 1.1% points greater than the difference before and after 1981, the break point in Table 1.

[6] Since this exercise is for illustrative purposes only, it is unnecessary to convert into real outlays. To do so would raise issues of how to account for different deflators for different components of federal outlays.

Chart B

Natural Logs of Federal Outlays, Its Defense, and Health/Medicare Components, 1956 to 2004
(Billions of current $)[7]



Source: author. For untransformed data: Table 1.1 Summary of Receipts, Outlays, and Surpluses or Deficits(-): 1789-2010, and Table 3.1 Outlays by Superfunction and Function, 1940-2004, U.S. Office of Management and Budget (2000) *Budget of the United States Government, Fiscal Year 2000* http://www.gpoaccess.gov/usbudget/fy00/pdf/hist.pdf. Accessed April 6, 2011.

---

[7] Scale Conversion for Chart B

Natural Log	Bn of Current $
7.5	1.8
10.0	23.0
12.5	270.0
15.0	3,250.0

The real issue is how to explain logarithms to a layperson who may have had an unhappy encounter with logarithms in high school? First, in the accompanying narrative, refer to the scale as proportional instead of logarithmic. Second, display the underlying data, not the logs, on the vertical axis as in Chart B.[8] Third, the narrative should be constructed in terms of ratios at given points in time, such as "according to Chart B, federal outlays for health in 1956 were only 1/4 of one percent of total federal outlays but by 2004 federal outlays for health, which now included Medicare, had risen by a factor of 60 to become 15% of total federal outlays."

Conclusions

Logarithmically transforming data has many advantages for trend analysis. It enables presentation of data that span several orders of magnitude on the same chart and it facilitates identification of event points where CAGR changes. However, logarithms are defined for positive numbers only; they cannot be used to represent zero or negative numbers.

[8] The underlying data is called the antilogs in the literature.

Appendix

Computing CAGRs

Deriving CAGR

[1] $Y_{t+1} = (1+a)Y_t$.

Here, **t** is a time index (0 = the present, 1 = one year from now, 2 = two years from now, etc.). The compound annual growth rate (CAGR) is the parameter **a**. In **n** years:

[2a] $Y_{t+n} = (1+a)^n Y_t$, or equivalently

[2b] $Y_{t+n} / Y_t = (1+a)^n$.

The formula for CAGR given in the introduction to this paper is derived from the latter equation by taking the n^{th} root of both sides and then subtracting one from both sides.

[3] $a = (Y_{t+n}/Y_t)^{(1/n)} - 1$

Slope of a Compound Growth Curve

To show that the slope of a compound growth curve using logarithmic data is a straight line, take the logarithm of both sides of equation [2b], assuming that all Y's are positive:

[4a] $\log(Y_{t+n}/Y_t) = n \log(1+a)$, or

[4b] $\log(Y_{t+n}) - \log(Y_t) = n \log(1+a)$

Dividing both sides of equation [4b] by **n** yields:

[5] $[\log(Y_{t+n}) - \log(Y_t)] / n = \log(1+a)$, which is constant.

The numerator of the left-hand side is the vertical separation between points associated with **t** and **t+n** and the denominator is their horizontal separation. This ratio is, by definition, the slope of a straight line between points at **t** and **t+n**. Since the right-hand side is constant, this proves that the slope of a compound growth curve, expressed logarithmically, is a straight line. The next page shows a demonstration of this proposition using Excel.

All statistical packages for computers, including Excel, compute the natural logarithm of any number and plot the results versus time. Alternatively, one does not need to compute logarithms if the data are plotted on semi-log graph paper with logarithms on the vertical axis. This tool is available on the Internet at http://www.math.neu.edu/fries/semilog-ex-hwk.pdf. One cycle corresponds to every order of magnitude in the data. The sample on the last page is 4-cycle semi-log paper with logarithms on the vertical axis.

1.000	2.50%
1.025	2.50%
1.051	2.50%
1.077	2.50%
1.104	2.50%
1.131	2.50%
1.160	2.50%
1.189	2.50%
1.218	2.50%
1.249	2.50%
1.280	2.50%
1.312	2.50%
1.345	2.50%
1.379	2.50%
1.413	2.50%
1.448	2.50%
1.485	2.50%
1.522	2.50%
1.560	2.50%
1.599	2.50%
1.639	2.50%
1.680	2.50%
1.722	2.50%
1.765	2.50%
1.809	2.50%
1.854	2.50%
1.900	2.50%
1.948	2.50%
1.996	2.50%
2.046	2.50%
2.098	2.50%
2.150	2.50%
2.204	2.50%
2.259	2.50%
2.315	2.50%
2.373	2.50%
2.433	2.50%
2.493	2.50%
2.556	2.50%
2.620	2.50%

to save space, intervening data are suppressed

7.210	2.50%
7.390	2.50%
7.575	2.50%
7.764	2.50%
7.958	2.50%
8.157	2.50%
8.361	2.50%
8.570	2.50%
8.784	2.50%
9.004	2.50%
9.229	2.50%
9.460	2.50%
9.696	2.50%
9.938	2.50%
10.187	2.50%
10.442	2.50%
10.703	2.50%
10.970	2.50%
11.244	2.50%
11.526	2.50%