SEPTEMBER 15. 1983

WHITE PAPER

Estimating Least Squares Growth Rates on Negative Data

BERKELEY RESEARCH GROUP

Prepared By:

Stuart A. McCrary, CFA | Director
SMcCrary@brg-expert.com
312.429.7902

Estimating Least Squares Growth
Rates on Negative Data

Introduction

The use of regression in investment analysis has become very common. One important model assumes that earnings (or some other time series) grow at some smoothed or averaged growth rate from a base year. The data and resulting regression parameters are used to estimate historical growth and forecast financial data into the future. Various measures of stability of earnings can also be calculated from the smoothed data.

Sometimes, the regression cannot be calculated. The regression is nonlinear and transformations cannot be applied to negative data. Also, the transformations may bias the results. The paper will study various ways of coping with negative data. Also, a method of solving the regression without resorting to log-transformations will be presented.

The first section will outline the problems in using log-linearly transformed data to estimate growth rates. Section II will discuss techniques in use for coping with negative earnings levels, Some weaknesses of these approximations will be presented. A solution to the growth rate estimation problem that does not involve the use of transformations is presented in Section III. Concluding remarks follow. The Appendix A lists a program written for the IBM Personal Computer to solve the problem. User documentation is provided in Appendix B.

Growth Rate Estimation from Log-linear Transformation

A constant growth model over T discrete time periods is presented in equation (1):

$$(1)\ E = C(1+g)^T$$

In the above equation, E (earnings, sales, cost, etc.) depends on only one variable, time (T). Two regression parameters are C – base year constant and g - least squares compound growth rate.

The problem is transformed to a linear equation by first taking the natural logarithm of each side:

$$(2)\ \ln(E) = \ln(C) + T * \ln(1+g)^T$$

Applying the following substitutions:

$Y = \ln(E)$

$X = T$

$a = \ln(C)$

$b = \ln(1+g)$

the equation becomes familiar:

$$(3)\ Y = a + bX$$

The model is easily solved by ordinary linear regression using one of many standard statistical packages or an inexpensive pocket calculator. Simply enter the values of X (time) and Y (log-transformed earnings) and find the regression parameters (a and b).

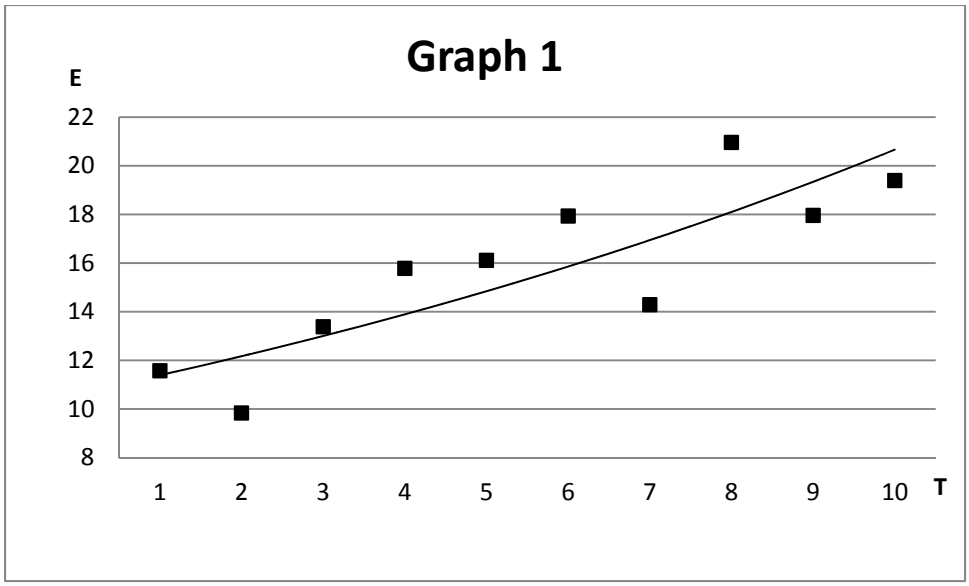
Logarithmic transformations make the solution quite easy. Unfortunately, the model cannot be used when the dependent variable (E) takes on negative values, because the natural logarithm of a negative number is not defined. It is common for earnings to be negative over an economic cycle. In these circumstances, the compound growth rate cannot be estimated with the method. Also, it is obvious that the model cannot be used when either of the regression parameters is negative (l+g or C). Naturally, when the regression cannot be calculated, the summary statistics such as R-Square are impossible to derive.[1]

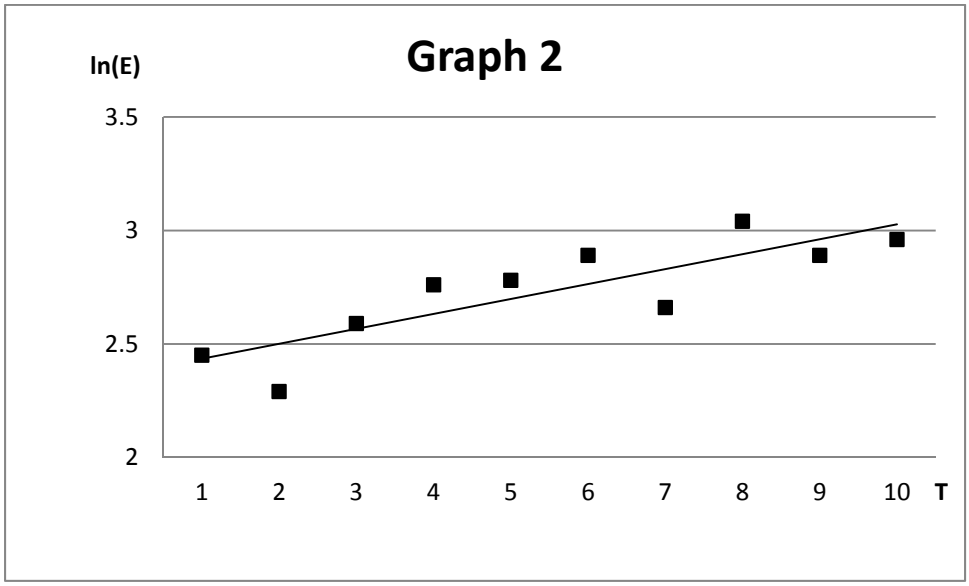
[1] R-Square measures the per cent of variation in the data explained by the regression. In general, a high R-Square is preferred to one with a low R-Square. Analysts using the least squares growth model to track earnings may use R-Square as a measure of risk.

A second problem arises with log-transformed regression. The logarithmic function is nonlinear. In fact, we must apply a nonlinear function to equation (1) to transform it into the linear equation (3). Below (Graph 1A) is a plot of a hypothetical set of data and regression curve fitted to equation (3).

Table 1 - Sample Data

T	E	ln(E)	
1	11.57	2.45	Regression Results
2	9.83	2.29	Using equation (3)
3	13.38	2.59	
4	15.78	2.76	a = 2.37
5	16.11	2.78	b = .066
6	17.93	2.89	g = 6.80
7	14.29	2.66	C = 10.69
8	20.96	3.04	R-Square = 71.3%
9	17.96	2.89	
10	19.39	2.96	





The error in the fit of the data is represented in the vertical distance (e) for each point. Fitting equation (1) minimizes the squared deviations in Graph 1.

The same data is log-transformed in Graph 2. Fitting equation (3) minimizes the squared deviations (e') represented in Graph 2. From the same fitted curve in Graphs 1 and 2, it is clear the transformation affects data at the far right of each graph much more than data at the far left. The effect is to compress the curve into a tighter range. As you would expect, the compression also affects the errors at the right of Graph 2 more than errors to the left. Since the far-right

errors are compressed, they carry relatively less weight in the regression. Ironically this data is usually the most recent data available. The log-transformation systematically gives more weight to old data and less weight to new data.[2]

Techniques to Handle Negative Data

Analysts have several techniques they use when they encounter negative data. The most common method simply omits the observations with negative values. Obviously, this technique introduces systematic bias in both regression parameters, because the smallest observations (the negative values) are excluded. Consequently, when the model is used to forecast earnings into the future, the forecasts will be upward biased. Measurements of error around the regression will also be biased. As a result, tests for goodness of fit, including R-Square, will be biased.

The amount of forecast error created by omitting data depends on many factors. A few omitted points in a very large sample will affect the regression statistics less than the same omissions from a small sample. Also, a larger number of omitted points will introduce more bias than a few excluded points, for a given sample size. The further the excluded points lie from the mean of the data, the greater the effect on the results. Because of these problems, omitting data is not a recommended strategy.

A second technique for coping with negative values is to add a constant. The constant must be as large as the absolute value of the most negative sample point. The technique can best be explained by an example. Suppose our sample followed a perfectly constant 10% growth rate as in Table 2A.

[2] The data in Table 1 generally follow a growth path with a positive growth rate (g). As a result, more recent data is generally larger than older data. If a time series follows a negative growth path, the pattern is reversed. For example, a regression of the number or cases of polio over time should reveal a pattern of reduction. A negative growth rate would measure the least squares reduction in the disease. For such data, log-linear transformations would still give more weight to some data than others. However, the most recent data would be emphasized.

Table 2A			Table 2B		
T	E		T	E	
1	10.00		1	110.00	
2	11.00		2	111.00	
3	12.10		3	112.10	
4	13.31		4	113.31	
5	14.64		5	114.64	
6	16.11		6	116.11	
7	17.72		7	117.72	
8	19.49		8	119.49	
9	21.44		9	121.44	
10	23.58		10	123.58	
11	25.94	(fitted)	11	124.35	(fitted)
12	28.53	(fitted)	12	125.97	(fitted)
13	31.38	(fitted)	13	127.60	(fitted)
14	34.52	(fitted)	14	129.25	(fitted)
15	37.97	(fitted)	15	130.92	(fitted)
a =	2.3		a =	4.68	
b =	0.1		b =	0.01	
C =	10		C =	107.94	
g =	10.00%		g =	1.30%	
R-Square =	100%		R-Square =	98.9%	

The log-linear regression correctly estimates g = 10%. The transformed model's R-Square indicates the model fits that data perfectly. Now suppose a constant of 100 is added to the data, as in table 2B. The log-linear least square growth rate is 1.30% with an R-Square of 98.9%. The R-Square is nearly identical in both regressions. The results in 2B could be used to estimate the R-Square in *2A.* It is not obvious, however, how to convert the least squares growth rate from Table 2B (1.30%) to the least squares growth rate in Table 2A (10.00%). Further, the tables give significantly different forecasted values for the next five years. Clearly, adding a constant is not an adequate approximation.

A third technique to find a log-linear growth rate when negative data are present uses a linear regression as an approximation. The regression of equation (3) is replaced by equation (4):

$$(4)\quad E = a' + b'T$$

In this model, the parameter b' is the average increase in E per period. The growth rate (g) is approximately the parameter b' divided by the average E.

$$(5)\quad g \cong b'/\overline{E}$$

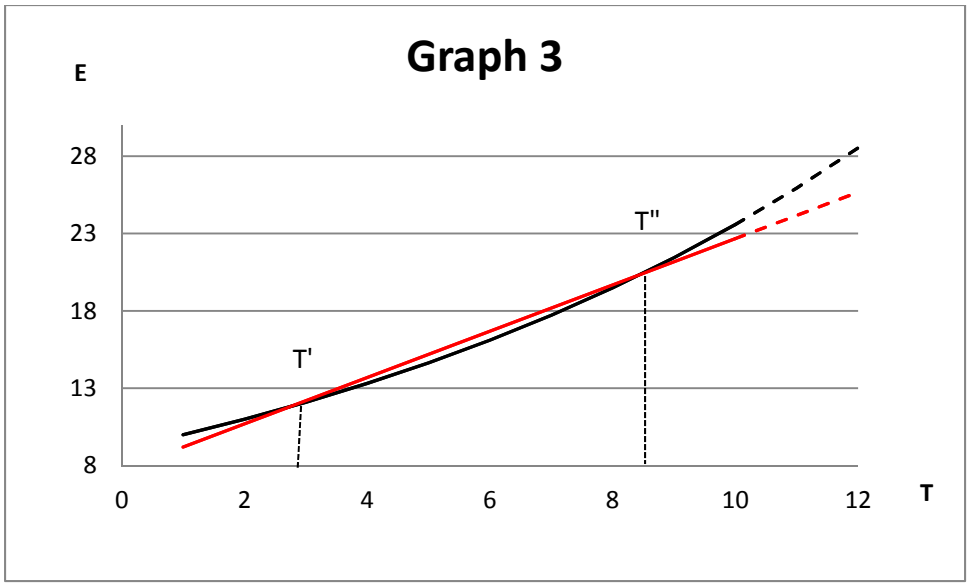
The R-Square of the model (4) is used as a proxy for R-Square of the model in equation (1).

The first weakness with this approach is that the error terms are systematic. That is, if the data series could most logically be viewed as a compound growth model, using another functional form will lead to errors in the regression that are not random. To see this, look to Graph 3. In the graph, the data from Table 2A has been plotted (a steady 10% growth rate). A linear regression as in equation (4) has been applied. The results are presented in Table 3.

Table 3

T	E-actual	E-fitted	Error
1	10	9.2	0.8
2	11	10.7	0.2
3	12.1	12.2	-0.1
4	13.31	13.69	-0.38
5	14.64	15.19	-0.55
6	16.1	16.68	-0.58
7	17.71	18.18	-0.47
8	19.49	19.68	-0.19
9	21.43	21.17	0.26
10	23.58	22.67	0.91

a' =	7.708
b' =	1.496
R-Square =	98.57%



As the chart shows, the linear model underestimated the smoothed value of E in the early and later stages of the model. The linear model overestimated E in the mid-years. Dotted line extensions represent forecasts. The chart clearly shows that linear regression quickly diverges from the log-linear curve.

It is also clear that the error statistics for the linear model will be different from the error statistics for the compound growth model. The data in Graph 3 was selected to perfectly fit the compound growth model, so it is not surprising that the R-Square in Table 3 is less than the R-Square in Table 2A. The use of a simple linear model cannot be rejected for all data. Particular data may in fact be better represented with a linear model than with equation (1). If the growth model is the best model for the problem, a linear approximation is a poor substitute.

One final point on the simple linear model must be added. The growth rate estimated in equation (5) is a simple growth rate. The growth rate in equation (1) is a compound growth rate. If the linear model must be used, the growth rate in equation (5) should be converted to a compound equivalent rate for comparability. A compound growth rate can be derived from the simple linear model via equation (6).

$$(6)\, g \cong \left(\frac{\text{Fitted E for last time (T) in the regression}}{\text{Fitted E for the first time (T) in the regression}}\right)^{1/\Delta/} - 1$$

This approximation is a constant compound growth rate between the most extreme points in the regression. Note that this g is an exact approximation if fitted values for T' and T" (Graph 3) are used because at these points, the linear model has no bias. These points T' and T" are not obtainable unless both models are solved (but if the compound growth model can be solved there is no need to approximate it) •

Direct Estimation of Regression Parameters

All the above techniques can be rather poor estimates of the problem presented in equation (1). The balance of this paper will address the problem directly and develop a solution without transformations. The model must be solvable even when negative data are present. The model must be solvable when negative regression parameters are expected.

The problem is to find the values for g and C that minimize the sum of squared errors in equation (1). The problem is stated as equation (7):

$$(7)\, Z = \sum \left(E - C(1+g)^T\right)^2$$

where Z is the sum of squared errors. The equation is minimized with respect to C and g where the partial derivatives of Z with respect to these parameters are equal to 0:

$$(8A)\, \delta Z/\delta C = -2\sum \left(E - C(1+g)^T\right) * (1+g)^T = 0$$

$$(8B)\, \delta Z/\delta g = -2\sum \left(E - C(1+g)^T\right) * CT(1+g)^{T-1} = 0$$

Equations (8A) and (8B) present two equations with two unknowns. The unknowns are the parameters that minimize the sum of squared errors. We solve for C in each of the above equations:

$$(9A)\,C = \frac{\sum E(1+g)^{T}}{\sum (1+g)^{2T}}$$

$$(9B)\,C = \frac{\sum ET(1+g)^{T-1}}{\sum T(1+g)^{2T-1}}$$

Combine the two equations:

$$(10) = \frac{\sum ET(1+g)^{T-1}}{\sum T(1+g)^{2T-1}} = \frac{\sum E(1+g)^{T}}{\sum (1+g)^{2T}}$$

Rewrite equation (10):

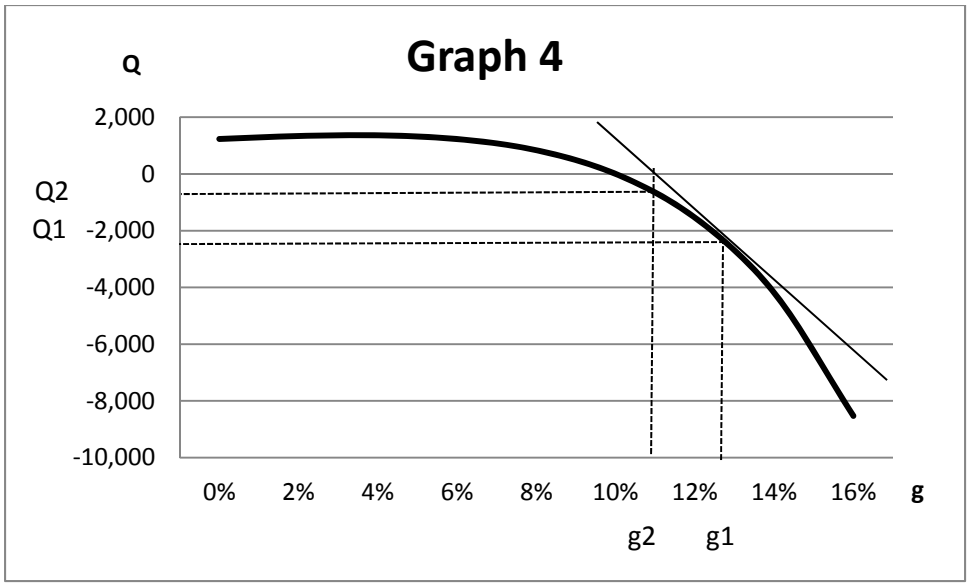
$$(11) = \sum ET(1+g)^{T-1} * \sum (1+g)^{2T} = \sum E(1+g)^{T} * \sum T(1+g)^{2T-1}$$

We must find the value (or values) of g that make this expression true. The value will not be found by algebraic solution. Rather, we will search for it using a Newton-Raphson search, a trial and error method of guessing the solution. Equation (11) will not be true unless the guess for g is the solution to the regression model. Equation (12) measures the error (Q) in equation (11) that results from choosing the wrong value for g.

$$(12)\,Q = \sum ET(1+g)^{T-1} * \sum (1+g)^{2T} - \sum E(1+g)^{T} * \sum T(1+g)^{2T-1}$$

The value for g that makes equation (11) true is the value for g that sets Q equal to O.

The technique is presented graphically in Graph 4.



An initial guess of g is to make g=g1. The error is Q1. The slope of the error function is calculated and used to estimate a new growth rate, g2. A new Q2 would be calculated, a new slope would be calculated and the process repeated.

The slope of the error function (equation 12) is the derivative of equation (12):

$$(13)\frac{dQ}{dg} = \sum ET(T-1)(1+g)^{T-2} * \sum (1+g)^{2T}$$
$$+ \sum 2T(1+g)^{2T-1} * \sum ET(1+g)^{T-1}$$
$$- \sum ET(1+g)^{T-1} * \sum T(1+g)^{2T-1}$$
$$- \sum T(2T-1)(1+g)^{2T-2} * \sum E(1+g)^{T}$$

The Newton technique is repeated until Q (the error in the sum of squares) is very close to O. Usually, the technique converges very quickly on the proper solution. Once g is known, C is readily calculated from 9A or 9B. The technique works for negative and positive data. Also, the regression is not biased by logarithmic techniques.

The regression parameters calculated above can then be used to refit the data. From the fitted data, the R-Square can be calculated.

Conclusion

Techniques for financial analysis have advanced rapidly in recent years. The introduction of the least squares growth model is an important step toward better financial analysis. The logarithm-based solution to the regression has several significant drawbacks and cannot be used if negative data are present. Various methods of overcoming this limitation are inadequate. The solution technique presented above does not require the user to exclude part of the data, it does not manipulate the data in misleading ways, and does not require the user to settle for inadequate linear estimates of the variables being measured.

Author's Note

September 5, 2013

A Newton search relying on Equation (12) and Equation (13) may sometimes fail to converge on the proper growth rate. The chance of convergence improves if the initial guess is greater than the solution value. Since the solution is not known when picking an initial guess, the author starts at a growth rate that is likely to exceed the solution value.

The best fit growth rate can also be found by bisection. Start with rates that are too high and too low and converge on the growth rate that sets Equation (12) to zero. Then use Equation (9A) or Equation (9B) to find C.

About Berkeley Research Group

Berkeley Research Group, LLC (www.brg-expert.com) is a leading global expert services and consulting firm that provides independent expert testimony, authoritative studies, strategic advice, data analytics, and regulatory and dispute support to Fortune 500 corporations, government agencies, major law firms, and regulatory bodies around the world. BRG experts and consultants combine intellectual rigor with practical, real-world experience and an in-depth understanding of industries and markets. Their expertise spans economics and finance, data analytics and statistics, and public policy in many of the major sectors of our economy, including healthcare, banking, information technology, energy, construction, and real estate. BRG is headquartered in Emeryville, California, with 25 offices in the United States, Australia, Canada, Latin America, and London, UK.