

---

## Return on Investment

**D**ata analysis is fully capable of solving all problems. It is incumbent, therefore, that the skilled analyst choose problems of the highest moral, ethical and scientific value. You can not squander your skills on anything less. Occasionally, however, it is necessary to teach the stuff and demonstrate its power by tackling showy, if less important matters — like getting fabulously rich. I'll start the exercise. Completion is an exercise for the reader.

The data are provided in Fortune's annual report of statistics for the "Fortune 500". I intend to find the variables that predicted the return I can get from investing my money. I will invest my money accordingly and then relax for a few years while my money multiplies. Then I will relax a great deal more.

First, a look at the variables. General Motors' data is shown but for each company the data include

*Revenues (Rank in 1995 and 1994)*

*The name of the company,*

*Revenues (In millions of dollars and as the percent by which revenues changed as compared to 1994.*

*Profit (in millions of dollars) and as change compared to 1994*

*Assets (in millions of dollars for 1996 and as an annualized (\*\*?\*of increase\*\*\*) during the preceding decade.*

*Total Stockholders' Equity (in millions of dollars)*

*Market Value (in millions of dollars)*

*Profits, shown as  
percent of sales  
percent of assets  
and as Percent of Stockholders' Equity*

*Earnings per share  
in dollars  
as percent increase compared to 1994  
and as an annualized rate of increase for the decade.*

*Total Return to Investors as  
percentage for 1995 and as  
annualized percentage for the decade*

*The number of employees*

*The Industry*

*And, finally, the address, and the name of the CEO.*

First, to preview the wealth I am about to attain: How well could I have done in 1995? To take a look,, I rank order everything according to percentage return for 1995 and I see .... and I see that 10% of the data are outright missing, with no value at all for this variable. So I begin my checklist of things that have to be checked before I can know what to make of any patterns I may find in the 90% of the data that are present.

Now, looking at the numbers, in rank order:

COMPANY	TOTAL RETURN TO INVESTORS			
			1985-95	
	1995		annual rate	
Name	%	FN	%	FN
Continental Airlines	353.3		—	
Northwest Airlines	223.8		—	
USAir Group	211.8		(8.9)	
Sun Microsystems	157.0		—	
Case	114.3		—	
Student Loan Marketing Assn.	109.5		7.7	
First Interstate Bancorp	108.5		15.6	
CompUSA	107.5		—	
UAL	104.3		13.9	
Seagate Technology	97.9		20.7	

Kmart	(42.3)	(0.8)
Jefferson Smurfit	(44.1)	—
Merisel	(45.3)	—
Yellow	(46.9)	(5.5)
Best Buy	(48.0)	17.6
Payless Cashways	(54.1)	—
Flagstar	(55.4)	—
Penn Traffic	(60.5)	—
Morrison Knudsen	(64.2)	(12.5)
Caldor	(85.4)	—

I see 353.3 percent for Continental Airlines, not bad, I can accept that rate of return. The top ten takes me down to a mere 97.9 percent, still not bad. And I notice a very promising regularity in the names: Continental Airlines, Northwest Airlines, USAir Group, and UAL. Unfortunately, in the cases where there is comparable data for the decade, the rates are not at all this high. Some (in parentheses) are negative. I also note that there is another end to this distribution, Caldor at the extreme (negative 85.4%). Perhaps some caution is in order.

This will have its own problems, but for the beginning of this project I think I will switch to returns annualized over the decade.

Now I see a lot more missing data, 119 of the "500", and looking at the top and

bottom of the list (of those for which there are data)

COMPANY	TOTAL RETURN TO INVESTORS	TO INDUSTRY	
		1985-95	
	1995	annual rate	
Name	%	%	
Home Depot	4.3	44.5	Specialist retailers
Conseco	45.8	42.6	Ins: life & health (stock)
Applied Materials	86.4	40.7	Electronics, electrical equipment
United HealthCare	45.0	38.3	Health care
Micron Technology	80.3	37.2	Electronics, electrical equipment
Nike	88.8	36.8	Wholesalers
Compaq Computer	21.5	36.1	Computers, office equipment
Computer Associates Intl.	76.5	35.1	Computer and data services
Fed. Natl. Mortgage Assn.	75.0	33.1	Diversified financials
Gillette	41.1	31.2	Metal products
Beverly Enterprises	(26.1)	(4.9)	Health care
Yellow	(46.9)	(5.5)	Trucking
Advanced Micro Devices	(33.6)	(5.5)	Electronics, electrical equipment
AST Research	(41.9)	(5.8)	Computers, office equipment
Turner Corp.	1.5	(7.3)	Engineering, construction
PriceCostco 11	18.4	(7.6)	Specialist retailers
USAir Group	211.8	(8.9)	Airlines
Unisys	(36.2)	(10.9)	Computer and data services
Morrison Knudsen 39	(64.2)	(12.5)	Engineering, construction
Navistar International6	(29.8)	(18.8)	Motor vehicles and parts

That is more believable, 30 to 40% per year, slightly more than two years to double (excluding taxes). Still not bad. There is also some greater reliability of prediction from decade statistics to year statistics, more than the other way around: The top 10 for the decade did well for the year (but not the other way around) — among those companies that survived the decade (and made it to the list). The list shows computing or electronics and health care among both the big winners and big losers.

Now, is the annualized rate of return a well behaved variable? This is a little troublesome. First I'm worried because this is certainly not a random sample of corporations, this is the Fortune "500", the largest (by assets). Moreover, the worst of the lot have probably disappeared, at least from the Fortune 500 and some of the worst and some of the best will have been acquired by others on the list. Whatever that did to the companies, it might well have affected the statistics. In addition, just dealing with percentages is troublesome. The intervals of a percentage imply that -20%, -10%, 0%, +10%, +20%, +30%, etc. are equal intervals. But percentages stand for simplified ratios. And as ratios, the ratio of .8 to .9 is the same as the ratio of .9 is to 1.01. As ratios, the equal steps are .8, .9, 1.01, 1.14, 1.28, 1.44.

## Equal steps as percentages

.8 .9 1.0 1.1 1.2 1.3

## Equal steps as ratios

.8 .9 1.01 1.14 .128 1.44

So what are the equal steps, even among percentages? This is a mess, particularly when I know that these particular ratios are annualized ratios out of a ten year span, which means that they have been treated as ratios and then, only for presentation, converted to percentages. So, if what I want is to represent data computed as ratios, and to represent them in a form that represents equal ratios as equal intervals, then I will use the logs of the ratios.

I like that argument and, with confidence I was about to show off by “discovering” that these numbers were well-behaved once converted to logs. Trouble is, they are not. In logs these returns have a skewed distribution with a tail to the left. (The sequence of mid values, from the median value to the mid thirty-second value, decreases.)

Count	Value	Value	Mid Value	
n=381				
191	0.130	0.130	0.130	Median
96	0.159	0.094	0.127	Mid Quartile
49.5	0.188	0.057	0.123	Mid Eighth
25	0.225	0.014	0.119	Mid Sixteenth

13	0.250	-0.028	0.111	Mid Thirty-Second
7	0.308	-0.060	0.124	Mid Sixty-Fourth

Because it violates my naive expectations, this would be a bit disconcerting, except that all the other complications affecting this distribution suggest that I give it a little leeway — if not the logs, then something close.

So, back to the numbers. Square roots. Still skewed to the left.

Using the original numbers. Still skewed to the left. Now I've given it more than a little "leeway". Still not well behaved.

Count	Value	Value	Mid Value	
n=381				
191	1.139	1.139	1.139	Median
96	1.172	1.099	1.136	Mid Quartile
49.5	1.207	1.059	1.133	Mid Eighth
25	1.252	1.014	1.133	Mid Sixteenth
13	1.284	0.972	1.128	Mid Thirty-Second
7	1.361	0.942	1.152	Mid Sixty-Fourth

So blindly, at least for the moment, I have to ask just what it would take to make this distribution symmetrical. And this is going in exactly the opposite direction from what I expected,

not

1, .5, 0 (no transform, square root, log,, decreasing)

but

1, 2, 3 ... (no transform, square, cube, increasing)

Proceeding, it takes something like a cube to get the median, mid-quartile, and mid-eighth to line up.

Count	Value	Value	Mid Value	
n=381				
191	1.478	1.478	1.478	Median
96	1.610	1.327	1.469	Mid Quartile
49.5	1.756	1.188	1.472	Mid Eighth
25	1.963	1.043	1.503	Mid Sixteenth
13	2.117	0.918	1.518	Mid Thirty-Second
7	2.521	0.836	1.678	Mid Sixty-Fourth

That is interesting. It violates my naive expectation, quite sharply, so I have to ask why — and maybe learn something.

I don't take the detail seriously, not the third power, as compared to the 2.5th power or the 2nd power. But I do take the direction leading to the 2nd or 3rd power seriously. What does it mean?

Consider the opposite. Suppose that logs had worked. In dollars, that would mean that each additional dollar is easier. One dollar profit on one dollar is hard, one dollar profit on one thousand dollars is easy.

This is running in the opposite direction. It means that once you have made a 1% return, the next 1% is “harder”, and the next harder, or less likely. Going down it means that going down 1% is easier if you’ve already dropped a couple of percent. In stock market terms it means that going up is more likely to be slow, going down more likely to be sudden. Don’t trust this — because I’m reaching in my mind for models or “folklore” to match the clue provided by the data — but that is what I read as a strong possibility, based on the shape of the distribution. In value terms, when a value is dropping, the chances increase that it will drop precipitously (get out fast). When a value is rising, the chances that it will continue to rise diminish.

Consider the implication. There is no proof here, nothing more than these numbers. But consider the implication to see the issue raised by this peculiar distribution. There is a minor industry organized to provide information to investors, very often graphical information. People look at these things. And although “technical analysis” of the stock market gives names and numbers to patterns, first people look at the numbers. Then they try to formalize patterns that may, or may not, be present.

To “look” at these numbers, charts are frequently provided on two scales:

One is the price of the stock. The alternative is the logarithm of the price of the stock — done for all the common sense reasons that we have already discussed about logarithms — constant rates of change (straight lines) on such graphs correspond to constant rates of increase at continuously compounded rates. This evidence, suggests that the place to look for simplicity lies in quite another direction, the squares or cubes, not the logs. Again — that is a lot of guessing, and at least a bit of sloppy thinking. (Note I slipped from ratios, which are a measure of change, to dollars which are not.) I do not submit this to you, or to myself as a fact, as a conclusion, or even as an inference. I am already criticizing it, reversing it, and generally turning it around as I think about it. I'm letting you see the usually hidden process of reasoning and guessing. This is what follows from looking at the shape of this distribution, finding a surprise, and thinking about it.

Can I predict these returns? For that I need some correlation with other variables, preferably a strong correlation. I'll begin with assets. Is it true that as corporations get larger, their rates of growth must diminish? Is it true that to find large returns I have to look to middle size, or smaller corporations.

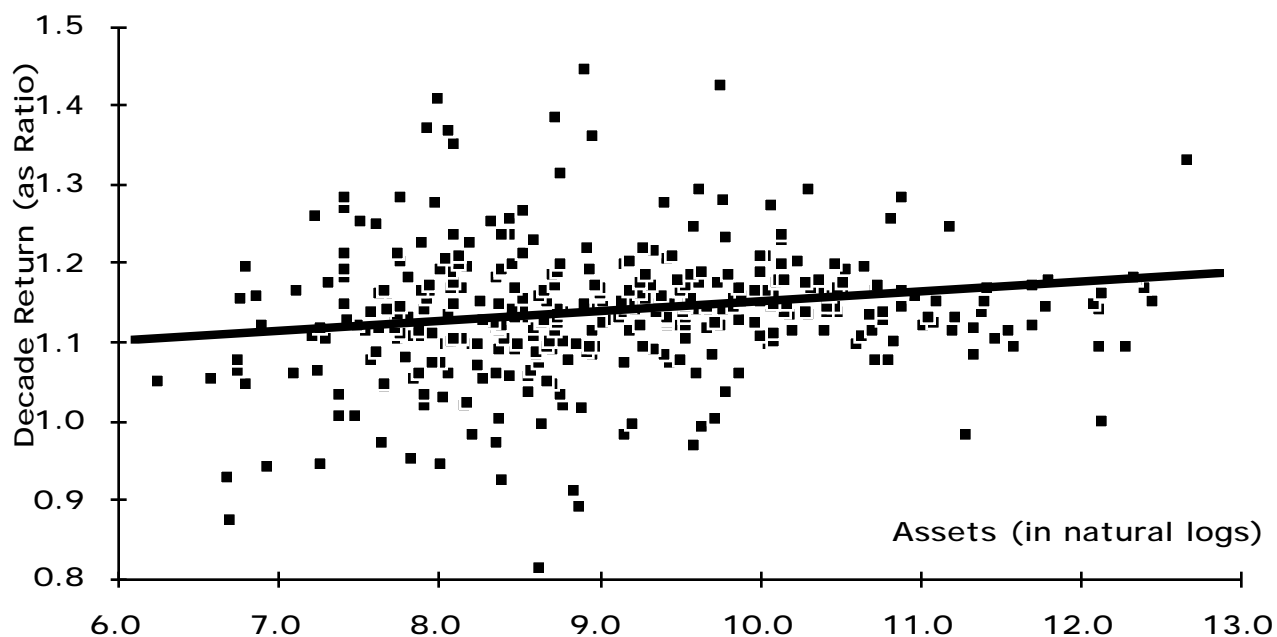
The assets variable is surely truncated — the selection criterion for

the Fortune list selected the largest corporations. So the first criterion for a well behaved variable can not be examined. Homeoscdasticity is a possibility. But I am going to do a quick start with logs and go directly to the relation, if any, between assets and return.

Using my spread sheet program, the number I get for the correlation is .14, and the number I get using the more exotic third power is .\_. That tells me not to worry a great deal about that transformation: Where most of the data lie between .95 and 1.3, curvature within this range is not going to be difficult to distinguish from a straight line.

Now that correlation tells me that approximately 2% of the variance in return is predictable from log assets. Can I be seriously interested in something that "explains" 2% of the variance in the variable I am trying to predict?

It all depends on context. And in this case I am very interested. Look at the graph of the data together with the regression line. It has an intercept of 1.0524 and a slope of .0092 (with a correlation of .145).



Just looking at the graph, there are lots of corporations to choose from with log assets approximately 7.5 (assets approximately 1.8 billion dollars). And there are lots of corporations to choose from with log assets approximately 10.5 (assets approximately 36.3 billion dollars). With a slope of .0092, the expected difference between their average returns (as ratios) is

$$.0092 \text{ times } 3 = .0276$$

That means that on the average the higher group realized 2.76% greater return. The expected return for the

higher group was 1.1491, as a ratio; the expected return for the lower group was 1.1215, as a ratio. Or, as percentages 15% versus 12% return on the investors' dollar. The variation around this return is enormous:

Reading it directly off the graph, it is common to find values between the predicted ratio  $-0.15$  and the predicted ratio  $+0.15$ .

Computing it :

The mean return, as a ratio, is 1.136, as a percentage that is 13.6% annual return to the investor.

The standard deviation of the ratio is .0805; the variance of the ratio is .0065.

In standardized form the variance of the residuals is  $\sqrt{10r\text{-squared}}$  is .9895, 98.95% unexplained.

Knowing that the variance of the return is .0065, that leaves the ratio with an error variance of .0064

and an error standard deviation of .0801, which is the square root of this number.

So, two standard deviations of error above and below corresponds to the predicted ratio  $\pm 1.602$  — very close to my eyeball estimate.

Can I find anything useful in such small correlations and large variance? You bet, literally, you *bet*. I can bet my money on 20 or 30 securities in the group with log assets of about 7.5 and get a return fairly close to the average for that group — and the average for that group is about 3% on the dollar greater than the lower asset group. Taking not those with assets of 7.5 in logs but taking the largest, and comparing their expected value to the mean return for all these investments, I should clear maybe 3% more than the average. And in the investment, beating the averages by 3%, is doing very well, gaining perhaps 16 or 17% per year rather than the average 13.5% per year.

So yes, this  $r$  is tiny. But for investment purposes my primary interest is in the slope. And the size of this slope is quite usable. Then, knowing that the  $r$  is small, I will have to protect myself by diversifying, in order to protect myself from the variance and realize the gains predicted by the slope.

So yes, I can work with these tiny correlations. The problem here is not the correlations. The problem here is

that I haven't *really* predicted anything. The words, "prediction", "estimation", "error", and so forth are statistical conventions. But check their meaning operationally, with these data: These numbers tell me how well I could have done, over the decade ending last year if, at the beginning of the decade, I had had the 1995 assets and the regression data that became available at the end of the decade. How well this variable will, or would, predict into the next decade or the next year is an interesting question, not answered by these data — though the exploration of these data provides an interesting lead.

#### Note on multiple regression

I have barely begun this exploration. Are their variables showing a higher correlation with return? (Yes)

It is also possible to try to predict one variable, Return, by writing a linear equation using two or more predictors, not just one.

$$y = a_0 + a_1x_1 + a_2x_2.$$

Most statistical programs will accommodate "multivariate regression". The interpretation of multivariate linear equations has some surprises in it, but the basics begin like the basics of two variable regression: Start with well behaved variables. Generate

residuals comparing  $y$  as predicted from the  $y$  that is observed. (you can graph these residuals as the ordinates in two or more different graphs of residuals, one for each of the predicting variables.) Then, if you want a number saying “how good”, compared the variance and standard deviation of the residuals to the variance and standard deviation of the original variable “ $y$ ”.