

Volume II: Lines

analyze ... to separate or break up (any whole) into its parts so as to find out their nature, proportion, function, relationship, etc.

— *Webster's New World Dictionary of the American Language, College Edition*

Lines

The prima donna of models in the social sciences is the simple straight line. Everyone knows what a line is: In a sketch it is just a trace on a graph — it has a direction and a certain height on the page. In geometry you learn that a straight line is determined by two points. In algebra you learn to match the geometry with an equation for the line, $y = mx + b$. In data analysis it says that the value of a variable, "y", is proportional to the value of a variable "x", with a constant "b" added to the result. And here it is: a straight line, drawn on graph paper and described with algebra.

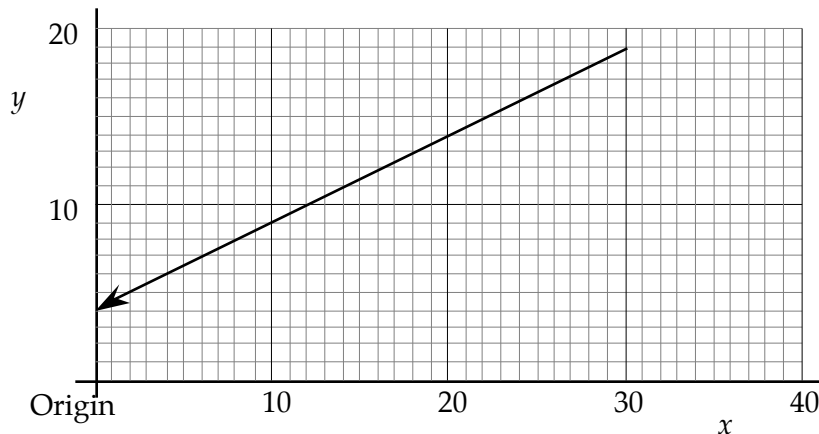


Figure __
A "mathematical" line, $y = mx + b$, $b = 4$, $m = .5$

The slope of the line is “ m ”, meaning that if the interval between two values of “ x ” is one, drawn horizontally, then the interval between the two corresponding values of “ y ” is m , drawn vertically. The intercept is “ b ”, meaning that if the value of x is 0 then the corresponding value of y is b . b is the height at which the line *intercepts* the vertical axis.

Using data, outside of mathematics, the line is our most-used way of fleshing out the detail by which one thing is related to another. If I describe the economic consequences of education by saying, “Comparing adult Americans, every year of formal education corresponds to an addition of three thousand dollars to the average income,” that description is a line with a slope of three thousand dollars per year. The intercept is unspecified by this description, but if it were zero it would mean that no education implied no money.

In much of the sciences, for descriptive work where theory is lacking, we have nothing comparable to the decidedly non-linear ellipses of planetary motion, the quadratics of acceleration, or oscillations of springs found in elementary theoretical physics. But that’s it — we use lines. On the other hand, using logs, and any other transformation that can help, we interpret “lines” so broadly that even the quadratic equations of velocity and acceleration can be handled *indirectly* by “linear” technique.

In data analysis it is rarely the case that we even have a continuous trace on a graph, straight or otherwise. In fact it is useful to amend one of the rules that mathematicians use for lines: In math, two things, the intercept and the slope, tell you everything there is to know about a line. By contrast, in data analysis, you need the intercept, the slope, and the scatter. Usually the best we have is a series of observations that line up (more or less)

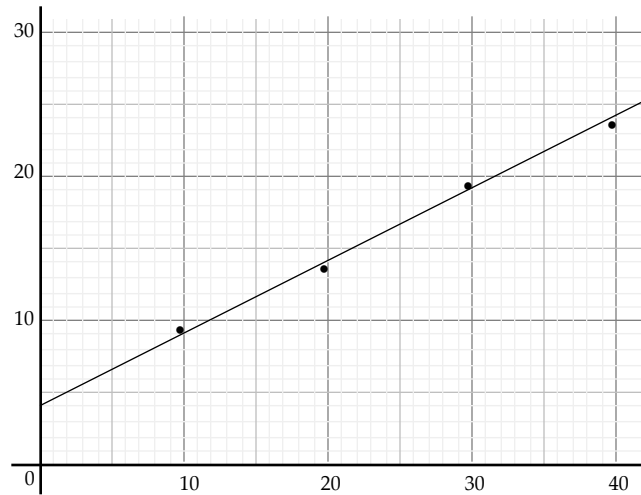


Figure A

And often what we have on the graph is less like a mathematically fine trace and more like an ambiguous cloud within which we may try to detect a pattern whose general shape may, possibly, and to some degree, be described by a line.

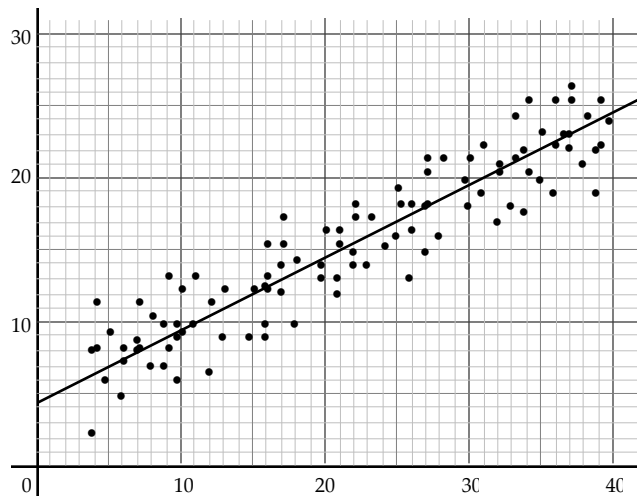


Figure B

It gets worse, much worse — the intercept, the slope and the “scatter”, usually a lot of scatter — leading to such aphorisms as “You only need strong statistics when you’ve got weak data.” If you are going to draw a graph describing 10,000 people, recording their numbers of years of education and their dollar incomes, then the folk wisdom — that there are people who never went to school, and got rich, while there are others who spent their lives in school, and made no money — will become a reality on your graph. You will have “points” (representing people) spread across all possible combinations of education and income. And it does no good to protest that the millionaire dropout and the impoverished Ph.D. don’t count. It does no good to protest that these are errors, or exceptions, or deviations. Such protests are like the complaints of biology students doing their first dissection of a frog: “Things aren’t where they are ‘supposed to be’”. No, things *are*. There is no *supposed to be* — there is nothing “defective” about the real world. Abstractions and averages are extremely valuable, indispensable really. But, no, sorry, the data are the reality. Reality is sovereign, not ideas. — This is what makes the line of the data analyst different from the line of the mathematician.

Beginning at the Beginning

Data analysis can get very complicated, at least as complicated as the world we try to discover through the analysis. That means you must proceed with caution. Simply launching into the data, drawing graphs, estimating lines, looking for correlations, letting your computer show off all the options of which your software is capable is no way to begin. You will simply overwhelm yourself with the possibilities, generate a mess, and quite possibly deceive yourself — either by thinking you have found things that are not there or, more likely, by failing to find things that are.

So for “multivariate analysis”, begin at the beginning or, as suggested for one variable analysis, begin before the beginning with “Who, What, Where, Why, When, and How?”.

Then the beginning for two variable is one variable analysis. These variables are the building blocks for two variable analysis with and they've got to be right — if they're not right you needn't bother with the rest. So, patience: Stem and leaf, looking for tell-tale patterns, looking for outliers and above all, looking for well-behaved variables.

Then, and only then, two variable analysis. And two variable analysis begins with a well-labeled graph. Very few computer software packages will give you a well-labeled graph. By well-labeled graph I mean something more than neat and pretty. By a well-labeled graph for two variables I am suggesting something comparable to the stem and leaf diagram for a single variable. It should organize the data so that my eyeball can look and so that my intuition can see. And what I am looking for is patterns. The graphs are so important that if I had to bet on results from two different analysts, one with a computer and conventional but inadequate software, the other tackling the data with graph paper, a pen, a ruler, and hand computation — I would bet on the second analyst.

Graphs

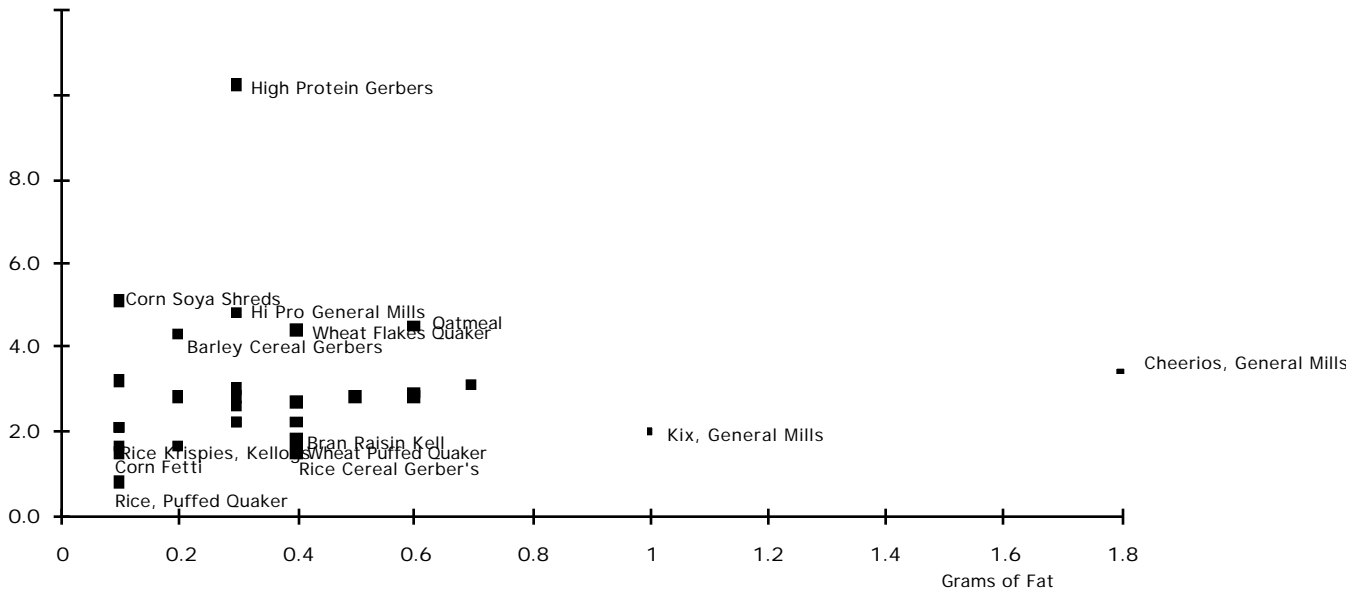
Figure 1 shows a two variable relation extracted from the breakfast cereal data. Beginning with each of the two variables, for a one variable analysis, each is close enough to being symmetrical. For Fat the median is 0.3 grams, the mid quartile is lower at 0.25 grams, and the mid-eighth is higher at 0.35 grams. For Protein the median is 2.75 grams, the mid quartile is lower at 2.65 grams, and the mid eighth is higher at 3.025 grams.

So, having completed the preliminaries I can proceed to the graph, What I see is outliers that are immediately apparent. I recognize the high protein cereal from previous work. I see Cheerios and Kix which sets me thinking about the bran component of cereals. I will forego the pleasures of another detailed analysis of breakfast cereals, except for two points.

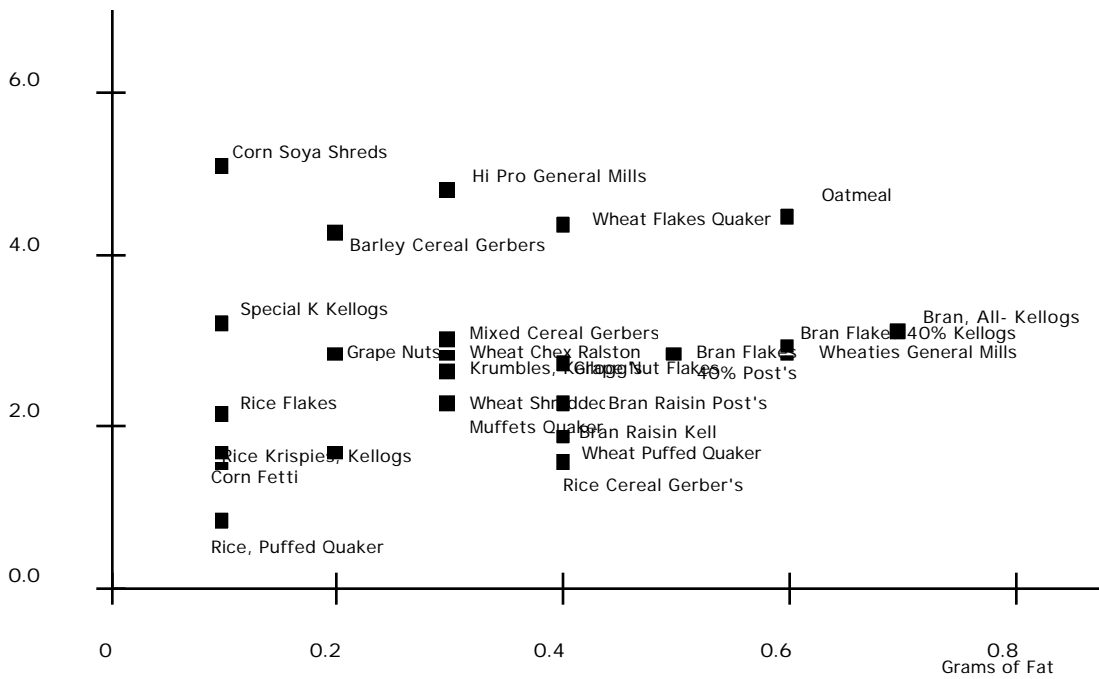
The first point is the labels: The labels on the points are the ones I think of when I refer to a well-labelled graph. As they did in a stem and leaf, these labels feed my intuition by allowing me to connect the data to other knowledge, other knowledge carried by the label itself, other knowledge that I have in my own experience and can connect to those labels. It is, I'll admit, a "pain" to draw such labels and even here I had to draw a second graph, a close-up, to get it done. It is difficult here with only 30 points. It is very difficult with more data points. But I have no need to be mechanical or fair to these data — I'm preparing the data to feed my intuition — so I will label outliers, I will label a few familiar points, I will label points that correspond to my hunches — pursuing the possibilities as is my purpose at the beginning of the analysis.

The second point is the relation itself: This pattern, overall, is hardly what I would call a line and that is an interesting result. It may be a weak correlation but that is very useful information. It is useful to know that the two nutrients occur relatively independent of one another.

Grams of Protein



Grams of Protein



Foregoing a deep and detailed analysis of breakfast cereals, in this case the initial two variable picture tells me little that is not present in the one variable information. The two variable picture tells me to go back to the one variable analyses: Learn what is to be learned about protein. Learn what is to be learned about fat (perhaps related to bran). The two discussions can be conducted independently, at least until they are better understood.