DM STAT-1 CONSULTING **BRUCE RATNER, PhD**

574 Flanders Drive North Woodmere, NY 11581 br@dmstat1.com
516.791.3544 ~ fax 516.791.5075 www.dmstat1.com

The Importance of Straight Data: Simplicity and Desirability for Good Model Building Practice

Bruce Ratner, Ph.D.

The purpose of this article is to show the importance of straight data for the simplicity and desirability it brings for good model building practice. I illustrate the topic sentence by giving details of what to do when an observed relationship between two variables depicted in a scatterplot is masking an acute underlying relationship. Data mining is employed to unmask and straighten the obtuse relationship. The correlation coefficient is used to quantify the strength of the exposed relationship, which possesses straight-line simplicity.

Straightness and Symmetry in Data

For DMers (statisticians, data analysts, data miners, knowledge discovers, and the like), exploratory data analysis, better known as EDA, places special importance on straight data, not in the least for the sake of simplicity itself. The paradigm of life is simplicity (at least for those of us who are older and wiser). In the physical world, Einstein uncovered one of life's ruling principles using only three letters: $E=mc^2$. In the visual world, however, simplicity is undervalued and overlooked. A smiley face is an unsophisticated, simple shape that nevertheless communicates effectively, clearly, and instantly. Why should DMers accept anything less than simplicity in their life's work? Numbers, as well, should communicate powerfully, unmistakably, and without more ado. DMers should seek two features that reflect simplicity – straightness and symmetry in data.

There are five reasons why it is important to straighten data:

1. The straight-line (linear) relationship between two continuous variables, say X and Y, *is as simple as it gets*. As X increases (decreases) in its values so does Y increase (decrease) in its values, in which case it is said that X and Y are positively correlated. Or, as X increases (decreases) in its values so does Y decrease (increase) in its values, in which case it is said that X and Y are negatively correlated. As an example of the above setting of simplicity (and everlasting importance) is Einstein's E and m have a perfect positive linear relationship.
2. With linear data, the data analyst without difficulty sees *what is going on within the data*. The class of linear data is the desirable element for good model building practice.
3. Most database models, belonging to the class of innumerable varieties of the statistical linear model, *require linear relationships* between a dependent variable and each predictor variable in a model, *and all* predictor variables jointly,

regarding them as an array of predictor variables that have a multivariate distribution.

4. It has been shown that *nonlinear models*, which are attributed with yielding good predictions with non-straight data, in fact *do better with straight data.*
5. I have not ignored the feature of symmetry. Not accidentally, as there are theoretical reasons, *symmetry and straightness go hand-in-hand*. Straightening data often makes data symmetric, and vice versa. Recall, symmetric data have values that are in correspondence in size, and shape on opposite sides of a dividing line or middle value of the data at hand. The iconic symmetric data profile in statistics is bell-shaped.

Data Mining is a High-concept

Data mining is a high-concept having elements of fast action in its development, glamour as it stirs the imagination for the unconventional, unexpected, and a mystic that appeals to a wide audience that knows curiosity feeds human thought. Conventional wisdom, in the DM space, has it that everyone knows *what data mining is*. [1] Everyone does it – that is what s/he says. I do not believe it. I know that everyone talks about it; but only a small, self-seeking group of data analysts genuinely does data mining. I make this bold and excessive self-confident assertion based on my consulting experience as a statistical modeler, data miner, and computer scientist for many years that have befall.

The Correlation Coefficient

The term *correlation coefficient*, denoted by **r**, was coined by Karl Pearson in 1896. This statistic, over a century old, is still going strong. It is one of the most used statistics, second to the mean. The correlation coefficient weaknesses and warnings of misuse are well documented. As a fifteen-year, plus or minus a few years, practiced consulting statistician, who also teaches statistical modeling and analysis, and data mining for continuing and professional studies for the Database Marketing/Data Mining Industry, I see too often the weaknesses and warnings are not heeded. Among the weaknesses and misuses, two are rarely mentioned: The correlation coefficient, whose values theoretically range within the left- and right-closed interval [-1, +1], is restricted in practice by the individual distributions of the two variables being correlated. [2] The misuse of the correlation coefficient is about its *linear assumption*, which is discussed below.

Assessing the relationship between a dependent variable and a predictor variable is an essential task in statistical linear regression, and nonlinear regression model building. If the relationship is linear, then test to determine whether the predictor variable has statistical importance to be included in the model. If the relationship is either nonlinear or indiscernible, then one or both of the two variables is/are reexpressed, or *data-mined* – being voguish with terminology, to reshape the observed relationship into a data-mined linear relationship. (I use interchangeably reexpressed, and data-mined.) Resultantly, the reexpressed variable(s) is/are tested for inclusion into the model.

The everyday method of assessing a relationship between two variables – lest the data analyst forgets: *linear* relationships, only – is based on the correlation coefficient. The correlation coefficient is often misused because its linearity assumption *is not tested*, albeit simple to do. (I put forth an obvious practical, but still not acceptable, reason why the non-testing has a long shelve life, later in the article.) I state the linear assumption, discuss the testing about the assumption, and provide how to interpret the correlation coefficient values:

> The correlation coefficient requires that the underlying relationship between two variables is linear. If the observed pattern displayed in the scatterplot of two variables has an outward aspect of being linear, then the correlation coefficient provides a reliable measure of the *linear* strength of the relationship. If the observed pattern is either nonlinear or indiscernible, then the correlation coefficient is inutile or offering risky results. If the latter data condition exists, data mining efforts should be attempted to straighten the relationship. In the remote situation when the proposed data mining method is not successful, then *extra*-data mining techniques, like binning, should be explored. The latter techniques are not discussed, because they are outside the scope of this article.
>
> If the relationship is deemed linear, then the *strength* of the relationship is quantified by an accompanying value of **r**. The following points are the accepted guidelines for interpreting the correlation coefficient:
>
> 1. 0 indicates no linear relationship.
> 2. +1 indicates a perfect positive linear relationship: as one variable increases in its values, the other variable also increases in its values via an exact linear rule.
> 3. -1 indicates a perfect negative linear relationship: as one variable increases in its values, the other variable decreases in its values via an exact linear rule.
> 4. Values between 0 and 0.3 (0 and -0.3) indicate a weak positive (negative) linear relationship via a shaky linear rule.
> 5. Values between 0.3 and 0.7 (0.3 and -0.7) indicate a moderate positive (negative) linear relationship via a fuzzy-firm linear rule.
> 6. Values between 0.7 and 1.0 (-0.7 and -1.0) indicate a strong positive (negative) linear relationship via a firm linear rule.

I present the sought-after scatterplot of paired variables (x, y) – a cloud of data points that indicates a *silver lining* of a straight-line. The correlation coefficient $\mathbf{r}_{(x, y)}$, corresponding to this scatterplot, assures that the value of **r** reliably reflects the strength of linear relationship between x and y. See Figure 1, below. The cloud of points in Figure 1 is not typical, due to the small, eleven-observation dataset used in the illustration. However, the discussion still holds true, as if the presentation involves, say, eleven thousand observations or greater. I take the freedom of writing to refer to the silver-lining scatterplot as a thin, wispy cirrus cloud.

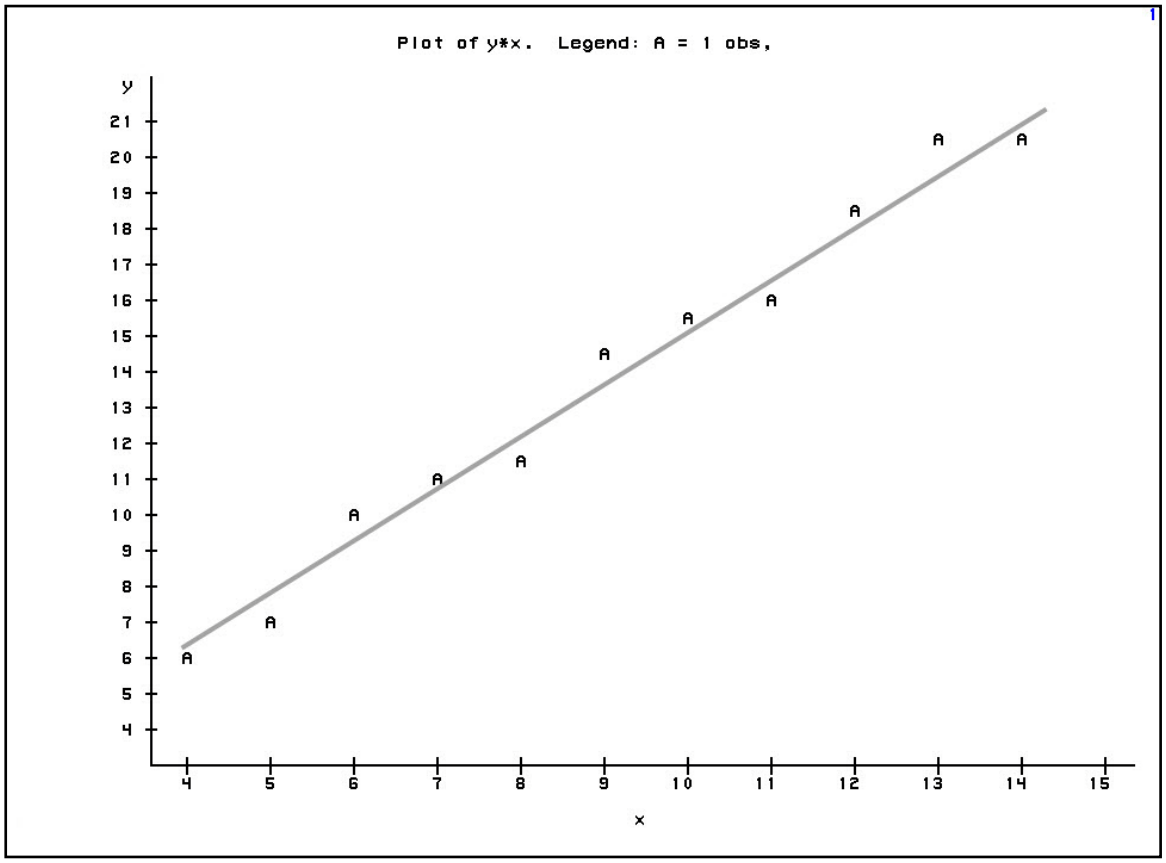

Figure 1. Sought-after Scatterplot of Pair (x, y)

Scatterplot of (x3, y3)

Consider the scatterplot of eleven data points of the third paired variables (x3, y3) in Table I, the well-known Ancombe Data, below. I construct the scatterplot of (x3, y3), in Figure 2, below, renaming (xx3, yy3) for a seemingly unnecessary inconvenience. (The reason for the renaming lies in the computer code of Table 3.) Clearly, the relationship between xx3 and yy3 is problematic: It would be straightaway linear if not for the *far-out* point ID#3, (13, 12.74). (Note: Far-out is not a voguish term for an outlier. See Tukey, J. W., *EDA*, Addison-Welsey, Reading, MA, 1977.) The scatterplot does not ocularly reflect a linear relationship. The nice and large value of $\mathbf{r_{(xx3,\ yy3)}=0.8163}$ is meaningless, and useless. I leave it to the reader to draw his/her underlying straight-line(!).

I. Ancombe Data

ID	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Anscombe, F. J., Graphs in statistical analysis, American Statistician, 27, 17-21, 1973.

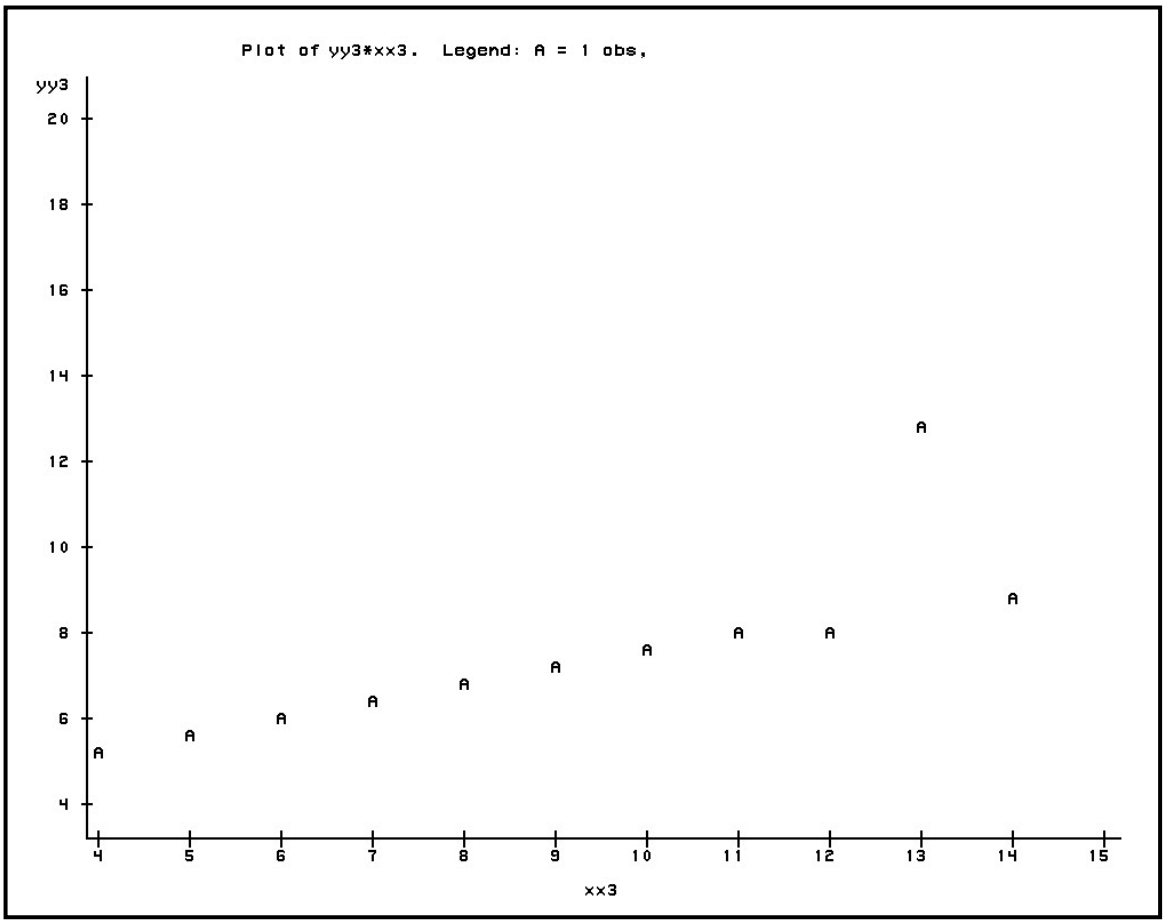

Figure 2. Scatterplot of (xx3, yy3)

Data Mining the Relationship of (xx3, yy3)

I data-mined for the underlying structure of the paired variables (xx3, yy3) using a machine learning approach under the discipline of evolutionary computation, specifically *genetic programming* (GP). The fruits of my data mining work yield the scatterplot in Figure 3, below. The data mining work is not an expenditure of preoccupied time (i.e., not waiting for time-consuming results) or mental effort, as the GP-based data mining (GP-DM) is a machine-intelligent, adaptively automatic process. The data mining software used is the GenIQ Model, which renames the data-mined variable with the prefix **GenIQvar**. Data-mined (xx3, yy3) is relabeled (xx3, GenIQvar(yy3)).

The correlation coefficient $\mathbf{r_{(xx3,\ GenIQvar(yy3))}=0.9895}$, and the uncurtained underlying relationship in Figure 3, warrants a silver, if not gold medal straight-line. The correlation coefficient is a reliable measure of the linear relationship between xx3 and GenIQvar(yy3).The almost maximum value of $\mathbf{r_{(xx3,\ GenIQvar(yy3))}}$ indicates an almost perfect linear relationship between the original xx3 and the data-mined GenIQvar(yy3). (Note: The scatterplot indicates GenIQvar_yy3 on the y-axis, not the correct notation GenIQvar(yy3) due to syntax restrictions of the graphics software.)

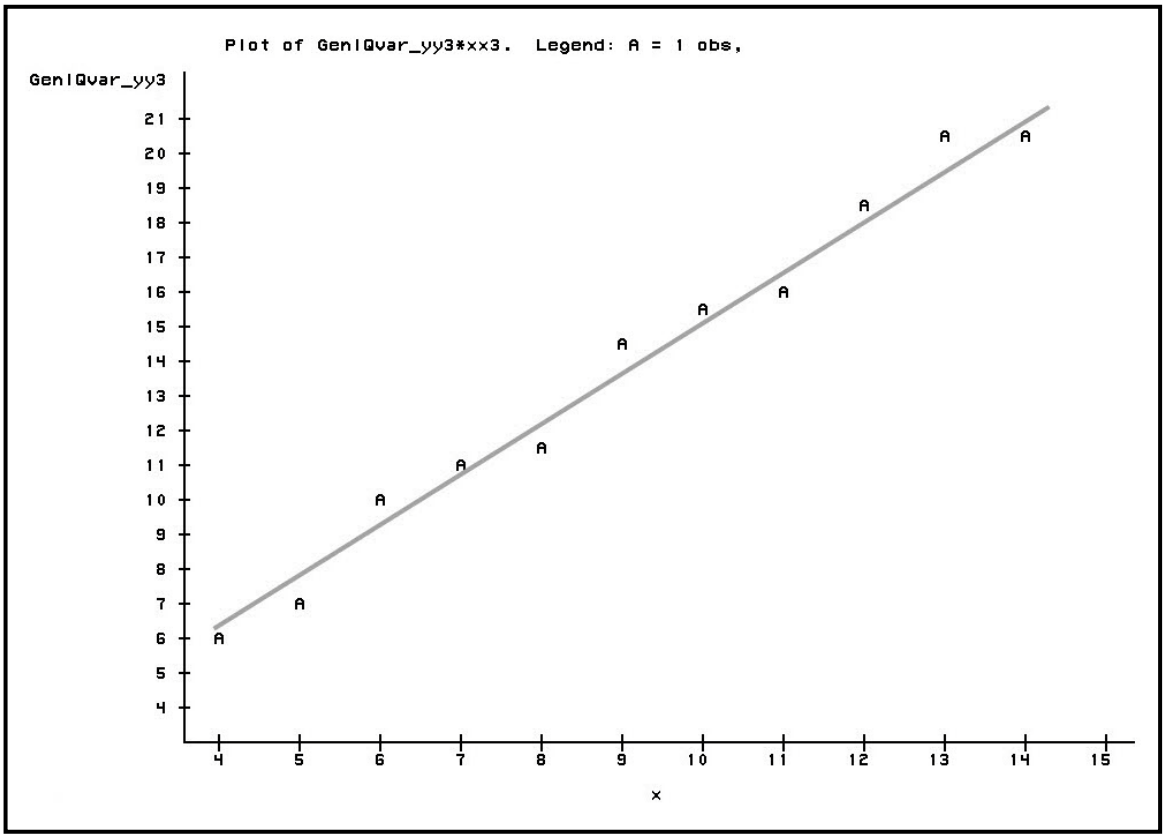

Figure 3. Scatterplot of (xx3, GenIQvar(yy3))

The values of the variables xx3, yy3, and GenIQvar(yy3) are in Table 2, below. The eleven data points are ordered based on the descending values of GenIQvar(yy3).

Table 2. Reexpressed yy3, GenIQ(yy3)
(Descendingly Ranked by GenIQ(yy3))

xx3	yy3	GenIQ(yy3)
13	12.74	20.4919
14	8.84	20.4089
12	8.15	18.7426
11	7.81	15.7920
10	7.46	15.6735
9	7.11	14.3992
8	6.77	11.2546
7	6.42	10.8225
6	6.08	10.0031
5	5.73	6.7936
4	5.39	5.9607

Side-by-Side Scatterplot

A side-by-side scatterplot of all the goings-on is best pictured, as it is worth a 1,000 words. The side-by-side scatterplot speaks for itself of a data-mining piece of work done well. See Figure 4, below.

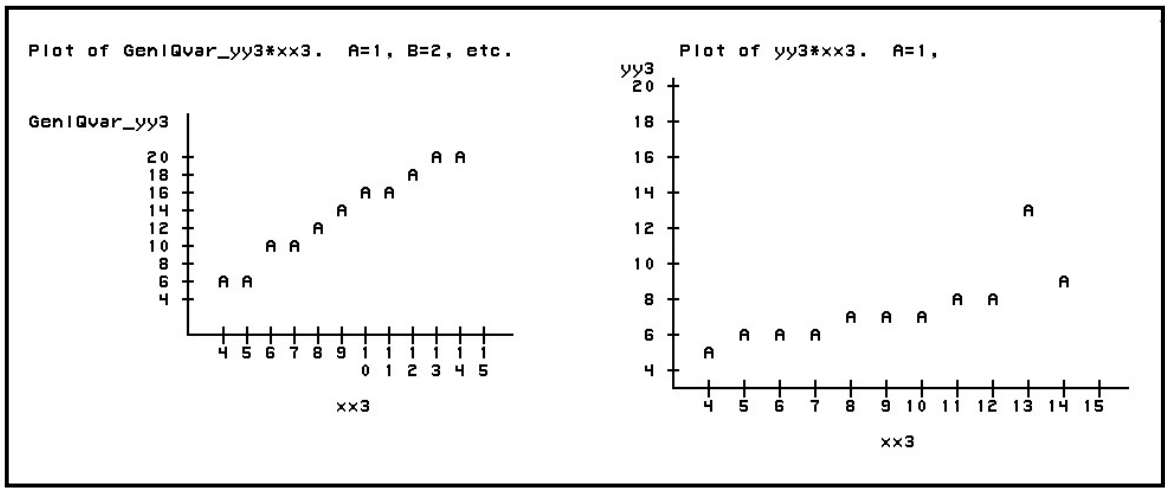

Figure 4. Side-by-Side Scatterplot

What is the GP-based Data Mining Doing to the Data?

As evidenced by the detailed illustration, GP-DM maximizes the straight-line correlation between a genetically reexpressed dependent variable and a single predictor variable. When building a multiple (many predictor variables) regression model, GP-DM maximizes the straight-line correlation between a genetically reexpressed dependent variable and each predictor variable, and the array of predictor variables jointly.

GP output is consists of two parts: 1) A *parse* tree, which provides some visual comfort of *what the model is doing.* The tree has the abstractness of a Picasso painting; it takes time to acquire an appreciation of the tree. 2) A corresponding *equation* (actually, a computer program/code), which in the present illustration is the wanted data-mined straightness within the eleven data points (x3, y3). (Note: Parse trees are use in the field of computer science, making appearances in computer programs, for which a given tree is a division of input into small sections that are easy for a program to process.) I present the GenIQ model, the tree and computer code, which did eminently good data mining that rendered the excellent, reliable result: $\mathbf{r_{(xx3,\ GenIQvar(yy3))}=0.9895}$.

The GenIQ Model Tree

The GenIQ Tree Display is in Figure 4, below.

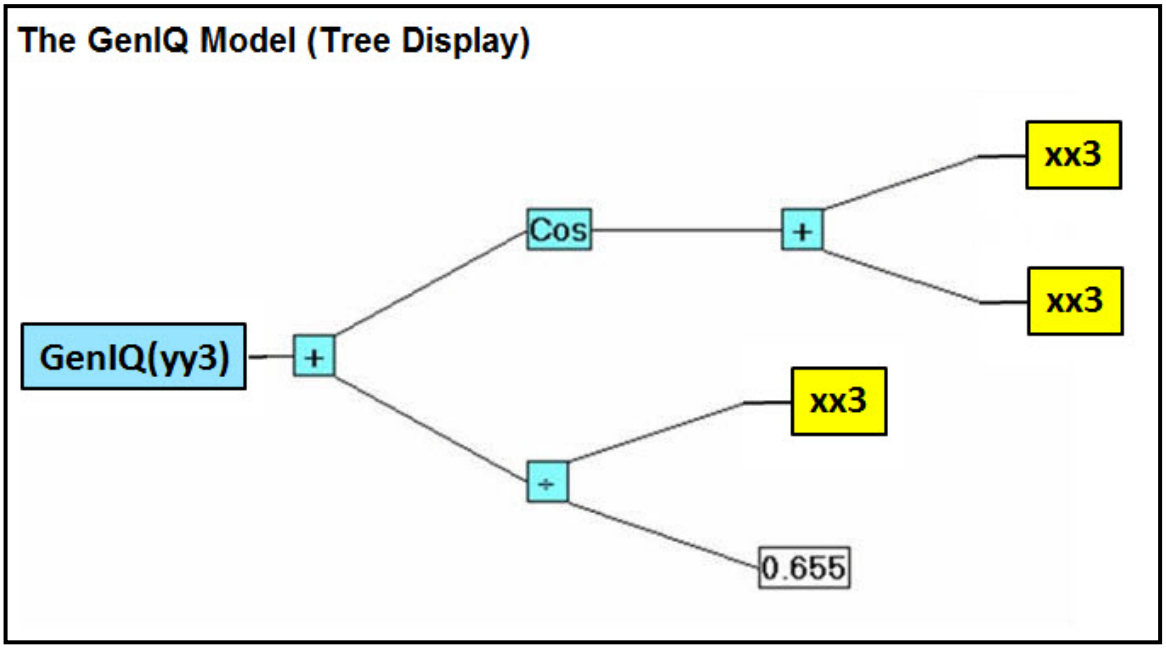

Figure 4. The GenIQ Model – Tree Display.

GenIQ Computer Code

The GenIQ Code is in Table 3, below.

Table 3. The GenIQ Model (Code)

```
x1 = .6550772;
        x2 = xx3;
   If x1 ne 0 then x1 = x2/x1;
        Else x1 = 1;
        x2 = xx3;
           x3 = xx3;
        x2 = x2 + x3;
        x2 = Cos(x2);
   x1 = x1 + x2;
GenIQvar(yy3) = x1;
```

Straightening a Handful of Variables, and a Dozen of two Baker's Dozens of Variables

A handful of variables (10 pairs of variables) can be dealt presumably without difficulty. As for a dozen of two baker's dozens of variables (703 pairs), they can be *handful*. However, GP-DM, as outlined with its main features and points of functioning, *reduces* initially, effectively, and efficiently the 703 pairs to a practical number of variables. This reduction operation is examined, below.

One requires clearness about a dozen of two baker's dozens of variables. A data analyst cannot expect to straighten 708 pairs of variables. (So many pairs require so many scatterplots. This is the reason for misusing, or ignoring the linearity assumption.) GP-DM can! In fact, it does so with the speed of a *gatling gun*. In practice, it is a common sight to work on a dataset of, say, 400 variables (7,980 pairs). GP-DM, in its beginning step, deletes variables (single and paired variables), which are deemed to have no predictive power by dint of the probabilistic-selected biological operators of reproduction, mating, and mutation. During the second evolutionary step, GP-DM further decreases the number of variables to only *handfuls.* The remaining step of GP-DM is the full-fledged evolutionary process of GP proper, by which the data strengthening is carried out in earnest.

Ergo, GP-DM can handle efficiently virtually any number of variables, as long as the computer being used has the capacity to process initially all the variables of the original dataset. Illustrating GP-DM with an enlarged dataset of many, many pairs of variables is beyond the sweep of this paper, too timbered for much required *paperspace*, not available here. Moreover, it would be of spurious injustice and reckless wronging of an exposition demanding, that which must go beyond 2D (two-dimensions), and even the 3D of the movie "Avatar."

Conclusion

The purpose of this article is to share to my personal encounters: Entering the mines of data, going deep to unearth acute underlying relationships, rising from the inner of the data mine – for showing the importance of straight data for the simplicity and desirability it brings for good model building practice. I discuss an illustration, simple but with an object lesson, of what to do when an observed relationship between two variables in a scatterplot is masking an acute underlying relationship. A GP-DM method is proposed directly toward unmasking and straightening the obtuse relationship. The correlation coefficient is used correctly when the exposed character of exampled relationship has straight-line simplicity.

Post Script

Any Question, Ask Bruce