DM STAT-I CONSULTING
BRUCE RATNER, PhD

574 Flanders Drive   North Woodmere, NY 11581
516.791.3544 fax 516.791.5075 1 800 DM STAT-1

br@dmstat1.com
www.dmstat1.com

A Genetic Programming Model:
Data-defined, Data Mining, Variable Selection, and Decile Optimization
Bruce Ratner, Ph.D.

The purpose of this article is to demonstrate the predictive power and features of a new genetic programming (GP) model – the GenIQ Model© – an alternative model to the statistical ordinary least squares and logistic regression models. The GenIQ Model, which is based on the assumption-free, nonparametric GP paradigm inspired by Darwin's Principle of Survival of the Fittest, offers theoretical and usable advantages over the two statistical regression models, which are long-standing, and widely used. [1, 2, 3] GenIQ automatically (requiring no programming despite the suggestive term programming in GP) *evolves* a model by letting the data define it: The GenIQ Model is data-defined. [4, 5] As well, GenIQ has four usable features, which are unique in the way they automatically and simultaneously begin and carry through to completion: 1) Data mine, 2) Variable selection, and 3) Set forth the model equation itself – such that, 4) The decile table, a measure of model performance, is optimized. [6, 7] The open-worked GenIQ Model and its wordbook are both generally regarded as not demanding on newcomers to GP modeling.

I use a small, real study using three variables (body fat percentage, age, and gender) to make GenIQ modeling tractable and attractive for the everyday modeler, with hope that GenIQ enjoys widespread and constant use. Addressing the study's first objective that requires a classification model, I build three classification models, a logistic regression model (LRM), a GenIQ Model, and a hybrid statistics-GenIQ Model, from which each model predicts the likelihood of a subject's gender is *male*. I illustrate GenIQs theoretical advantage via the GPP over the antithetical statistical paradigm – fitting data to a pre-specified model, which is defined by the rigorous methodology of significance testing, recently the focus of an ideological debate to abandon it. [8] Likewise, I illustrate that GenIQ variable selection provides more information and generates a better subset of predictor variables than statistical variable selection methods, which are viewed by many statisticians as suboptimal. [9] As for a GenIQ data-mining counterpart, no data mining capability for the two statistical regressions models exists. Certainly, I explain GenIQ output that consists of two parts – the *equation*, actually, a computer program, and a visual display of the program, often likened to a mathematical tree with Picasso-like abstractness – that account for the limited use of GenIQ. Admittedly, GenIQ is distinguished by its singular weakness of *difficulty* in interpreting the GP-based model.

There is a second study objective that I unparalleled address with GenIQ: How are AGE and PERCENT_FAT related? The latter objective points to nine extra-GenIQ applications. [10] I arbitrarily choose a LRM scenario for presenting GenIQ in an orderly, detailed way, but all that is presented and implied holds true for an ordinary least-squares regression model (OLS) scenario for presenting GenIQ.

Presenting the GenIQ Model as a viable alternative of LRM or OLS, I in effect put forward a *trinity of the regression paradigm*, from which modelers can now consider:

1) The GPP/GenIQ with an explicit fitness function for decile optimization; four features with unique execution; and GenIQs ungainly interpretability.

2) The Statistics paradigm/LRM/OLS with an unknowingly implied decile optimization achieved with LRM/OLS fitness functions (i.e., joint probability likelihood function, and mean squared error, respectively) serving as surrogates for explicit fitness functions for decile optimization; and the salient feature of model interpretability that is made possible by the regression coefficients.

3) The Hybrid Statistics-GP paradigm – integrating the best characteristics of two paradigms – yields a utile alternative of LRM/OLS or the GenIQ Model. The hybrid paradigm is: The modeler fits the data to LRM/OLS with the modeler's preferred variable selection method to determine the best subset among the original and GenIQ *genetically* data-mined variables. Of primary import, the hybrid LRM/OLS-GenIQ Model includes the regression coefficients, which provide the necessary comfort level of model interpretability for model acceptance.

I. <u>Situation</u>

The data are from a study investigating a new method of measuring body composition. I dub the study the *Phat Example*, which consist of three predictor variables: Body fat percentage (PERCENT_FAT), AGE, and gender/MALE (there are four males) for eighteen normal adults aged between 23 and 61 years. [11] The target variable is defined as: MALE=1 if subject is a male; MALE=0 if subject is a female. There are two objectives of the study: Is there any evidence that allows discrimination among males and females? If a model can be built to predict the likelihood of a subject's gender is male, then the model is the evidence. The second objective: How are AGE and PERCENT_FAT related? I first build a *MALE* classification model, and then explore the relationship between AGE and PERCENT_FAT. The Phat Example data are in Table 1, below. The table is arbitrarily ascendingly ranked by AGE, and the males are highlighted in green. A point of note: I highlight the males, and sometimes their accompanying variables, in all tabled results throughout the article.

Table 1. The Phat Example Data

ID	AGE	PERCENT_FAT	MALE
1	23	0.9	1
2	23	2.7	0
3	27	0.7	1
4	27	1.7	1
5	39	3.1	0
6	41	2.5	0
7	45	2.7	1
8	49	2.5	0
9	50	3.1	0
10	53	3.4	0
11	53	4.2	0
12	54	2.9	0
13	56	3.2	0
14	57	3	0
15	58	3.3	0
16	58	3.3	0
17	60	4.1	0
18	61	3.4	0

I build a LRM and a GenIQ Model, for the target variable MALE. The two classification models provide a counterpoint where the modeler can choose between the current standard interpretable LRM, and the new, potentially better, yet ungainly interpretable GenIQ Model. I make special mention of: An optimal decile table in the Phat Example is equivalent to the best ranking of the subjects, namely, all four males occupy the top four rank positions, based on the ascending LRM/GenIQ Model scores.

II. Phat-LRM Output
The LRM output (Analysis of Maximum Likelihood Estimates) – the Phat-LRM model/equation is:

Log of odds of MALE (=1) = 11.0912 + 0.00940*AGE - 4.9393*PERCENT_FAT

III. Phat-LRM Results

The results of the Phat-LRM are in Table 2, Rank Predictions of MALE based on Log_of_odds_of_MALE, below. There is not a perfect rank prediction of MALE, as male ID #7 is in the sixth rank position. Thus, I declare the Phat-LRM is a good model with adequate predictive power. (It is well understood that no validation has been performed.)

Table 2. Rank Prediction of MALE based on Log_of_odds_of_MALE

ID	AGE	PERCENT_FAT	MALE	Log_of_odds_of_MALE
3	27	0.7	1	7.88749
1	23	0.9	1	6.86203
4	27	1.7	1	2.94819
8	49	2.5	0	-0.79645
6	41	2.5	0	-0.87165
7	45	2.7	1	-1.82191
2	23	2.7	0	-2.02871
12	54	2.9	0	-2.72517
14	57	3	0	-3.1909
9	50	3.1	0	-3.75063
5	39	3.1	0	-3.85403
13	56	3.2	0	-4.18816
15	58	3.3	0	-4.66329
16	58	3.3	0	-4.66329
18	61	3.4	0	-5.12902
10	53	3.4	0	-5.20422
17	60	4.1	0	-8.59593
11	53	4.2	0	-9.15566

IV. Phat-GenIQ Model Output

The Phat-GenIQ Model output, which consists of a Tree Display and an Equation (Computer Program), is in Figures 1 and 2, respectively, below. The Phat-GenIQ Tree is compact, easy to read, but not easy to interpret. I build a second Phat-GenIQ Model, for reasons I discuss in section VIII. As a side benefit of the second model, I hope that *as exposure* to the GenIQ Model output increases, the modeler's comfortability of GenIQ modeling increases, which in turn, increases the acceptability of the GenIQ Model.

Figure 1. The GenIQ Model (Tree Display)
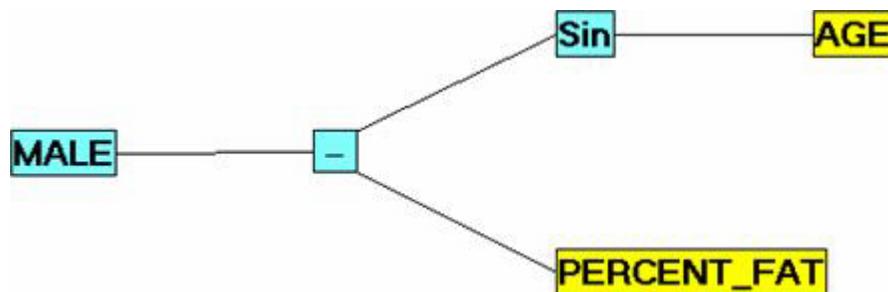
Figure 2. The GenIQ Model (Computer Program)

```
x1 = PERCENT_FAT;
    x2 = AGE;
        x2 = Sin(x2);
    x1 = x2 - x1;
GenIQvar = x1;
```

V. GenIQ Variable Selection

GenIQ variable selection provides a ranking of variable importance for a model's predictor variable (against a model's target variable), with respect to the other model's predictor variables *considered jointly*. Moreover, GenIQ ranks each remaining candidate predictor variable (not in the model), with respect to *all the variables* (in and out of the model) considered jointly. This is in stark contrast to the well-known, always-used correlation coefficient, which only provides a measure of linear strength of a predictor variable and a target variable, *independent* of all the variables. Additionally, it is believed that the regression coefficients provide a reliably listing of which model's predictor variables are most important, in rank order, accounting for the other predictor variables in the model. However, it is not well known that the standard method of

interpreting regression coefficients often leads to an incorrect ranking of variable importance for a predictor variable with respect to other predictor variables considered jointly. [12]

This study has two predictor variables, providing the simplest case to GenIQ ranking of variable importance. GenIQ identifies PERCENT_FAT as having more predictive power than AGE.

GenIQ Variable-Importance (with respect to all the variables considered jointly)

    1. PERCENT_FAT
    2. AGE

VI. GenIQ Data Mining

GenIQ data mining is directly apparent from the Phat-GenIQ Tree itself in Figure 1, above. The tree has two genetically data-mined structures (i.e., new variables, the spoils sought of a data mining effort): The sine transformation of AGE, sin(AGE), denoted by sine_of_AGE, and the Phat-GenIQ Tree itself, a single branch defined as sin(AGE) minus PERCENT_FAT. *Why* the sine function has an outward aspect in the Phat-GenIQ Model is beyond the scope of this article. But, suffice it to say, the GP modeler considers many functions beside trigonometric functions based on experience, and trial and error. *How* the sine function has an appearance in the model is due to GenIQ letting the data and functions define the model. In the second Phat-GenIQ Model, there is the *signature* of GenIQ data mining, namely, many branches, actually six. See section IX, or perhaps wait for the presentation of the second Phat-GenIQ!

VII. Phat-GenIQ Model Results

The Phat-GenIQ Model Results are in Table 3, Phat-GenIQ Model GenIQvar Rank Prediction of MALE, below. There is a perfect rank prediction of MALE based on Phat-GenIQ Model score GenIQvar. Thus, regarding the first objective: Is there any evidence that allows discrimination among males and females? Yes, the Phat-GenIQ Model provides a perfect ranking of the males, as they are all in the top four rank positions. The Phat-GenIQ Model is the evidence sought. As well, I declare the Phat-GenIQ Model an excellent model with perfect predictive power. (It is well understood that no validation has been performed.)

Table 3. Phat-GenIQ Model GenIQvar Rank Prediction of MALE

ID	AGE	PERCENT_FAT	MALE	GenIQvar
3	27	0.7	1	0.2563759
4	27	1.7	1	-0.743624
1	23	0.9	1	-1.74622
7	45	2.7	1	-1.849096
5	39	3.1	0	-2.136205
15	58	3.3	0	-2.307127
16	58	3.3	0	-2.307127
14	57	3	0	-2.563835
6	41	2.5	0	-2.658623
10	53	3.4	0	-3.004075
9	50	3.1	0	-3.362375
8	49	2.5	0	-3.453753
12	54	2.9	0	-3.458789
2	23	2.7	0	-3.54622
13	56	3.2	0	-3.721551
11	53	4.2	0	-3.804075
18	61	3.4	0	-4.366118
17	60	4.1	0	-4.404811

VIII. The Predictive Prowess of the GenIQ Model

To appreciate the predictive *prowess* of the GenIQ Model it is worthy to examine the individual relationship for each predictor variable and the target variable MALE. The observed relationships, indicating Rank Prediction of MALE based on AGE, sine_of_AGE, and PERCENT_FAT, are in Tables 4, 5 and 6, respectively, below:

1. AGE is a *fair* predictor of MALE with a correlation coefficient of -0.66. AGE is negatively related to MALE because the four males are in the lower bottom seven rows of Table 4, below. AGE is a fair predictor because the last row is female ID #2, and two males IDs #3 and #7 are between females IDs #5 and #6. The other two males IDs # 1 and #4 are above female ID # 2.

2. Sine_of_AGE is a *poor* predictor of MALE with a correlation coefficient of 0.28. Sine_of_AGE is positively related to MALE because a cluster of three males is in the top six rows of Table 5, below. Sine_of_AGE is a poor predictor because the four males are in the middle (twelve rows) of Table 5, despite the cluster of males IDs #3, #4 and #7. The latter males are in the bottom of upper six rows, and are followed by three females IDs #15, #16, and #5. The fourth male ID #1 is in the fourth row from the bottom row.

3. PERCENT_FAT is a *better-than-fair* predictor of MALE with a correlation coefficient of -0.78. PERCENT_FAT is negatively related to MALE because all the four males are in the bottom seven rows of Table 6. PERCENT_FAT is a better-than-fair predictor because three males IDs #4, #1 and #3 are in the bottom three rows. The fourth male ID #7, being in the sixth row from the bottom row, contributes to PERCENT_FATs predictiveness.

In sum, there are three wanting predictor variables with poor to better-than-fair predictor power, and inconsistent data elements among several subjects (discussed in section XI). But nevertheless, GenIQ shows its predictive prowess by evolving a model to produce a <u>perfect</u> rank prediction of MALE based on the Phat-GenIQ Model in Table 3, above. It is interesting to note that the Phat-GenIQ Model uses a poor variable, and a better-than-fair predictor variable to yield the best ranking.

Table 4.
Rank Prediction of
MALE based on AGE

ID	MALE	AGE
18	0	61
17	0	60
15	0	58
16	0	58
14	0	57
13	0	56
12	0	54
10	0	53
11	0	53
9	0	50
8	0	49
7	1	45
6	0	41
5	0	39
3	1	27
4	1	27
1	1	23
2	0	23

Table 5.
Rank Prediction of
MALE based on sine_of_AGE

ID	MALE	sine_of_AGE
15	0	0.9928726
16	0	0.9928726
5	0	0.9637954
3	1	0.9563759
4	1	0.9563759
7	1	0.8509035
14	0	0.4361648
10	0	0.3959252
11	0	0.3959252
6	0	-0.158623
9	0	-0.262375
17	0	-0.304811
13	0	-0.521551
12	0	-0.558789
1	1	-0.84622
2	0	-0.84622
8	0	-0.953753
18	0	-0.966118

Table 6.
Rank Predictions of
MALE based on PERCENT_FAT

| ID	MALE	PERCENT_FAT
11	0	4.2
17	0	4.1
10	0	3.4
18	0	3.4
15	0	3.3
16	0	3.3
13	0	3.2
5	0	3.1
9	0	3.1
14	0	3
12	0	2.9
2	0	2.7
7	1	2.7
6	0	2.5
8	0	2.5
4	1	1.7
1	1	0.9
3	1	0.7

IX. Phat-GenIQ Model Version #2 Output and Results

GenIQ modeling is like other (non-physical science) modeling methods: There is no unique solution for a given dataset. Alternative solutions come from alternative methods, or from different versions of the same method. Regarding the latter, I build a Phat-GenIQ Model Version #2. The Phat-GenIQ Model Version #2 Tree Display and Computer Program (which includes *Int*, the Integer function that takes the integer part of a number) are in Figures 3 and 4, below. The Phat-GenIQ Model Version #2 GenIQvar2 Rank Prediction of MALE is in Table 7, below. The Phat-GenIQ Model Version #2 produces a perfect rank prediction of MALE. The second Phat-GenIQ Model is additional evidence indicating that discrimination between males and females can be made. (It is well understood that no validation has been performed.)

For sake of completeness, I discuss GenIQ variable selection and data mining for the second Phat-GenIQ Model. GenIQ variable selection unsurprisingly has the same finding as the first Phat-GenIQ Model: PERCENT_FAT has more predictive power than AGE. However, GenIQ data mining for the second model is obviously not the same as the first model, which has only two genetically data-mined variables, discussed in section VI, above. The second Phat-GenIQ Model has six genetically data-mined variables:

1. New_var1 = AGE * 0.141
2. New_var2 = Int (AGE * 0.141)
3. New_var3 = Int (PERCENT_FAT)
4. New_var4 = New_var2 * New_var3
5. New_var5 = New_var4 / AGE
6. New_var6 = New_var5 / New_var2, this is the Phat-GenIQ Model Version #2

Figure 3. The GenIQ Model Version #2 (Tree Display)
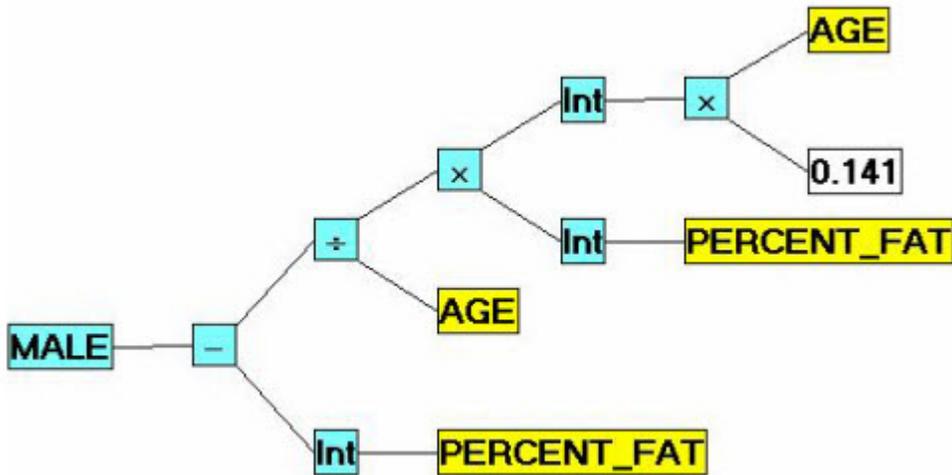
Figure 4. The GenIQ Model Version #2 (Computer Program)

```
x1 = PERCENT_FAT;
   x1 = Int(x1);
      x2 = AGE;
         x3 = PERCENT_FAT;
         x3 = Int(x3);
            x4 = .1407257;
               x5 = AGE;
            x4 = x4 * x5;
            x4 = Int(x4);
         x3 = x3 * x4;
   If x2 NE 0 Then x2 = x3 / x2; Else x2 = 1;
   x1 = x2 - x1;
GenIQvar2 = x1;
```

Table 7. Phat-GenIQ Model Version #2 GenIQvar2 Rank Prediction of MALE

ID	AGE	PERCENT_FAT	MALE	GenIQvar2
1	23	0.9	1	0
3	27	0.7	1	0
4	27	1.7	1	-0.888889
7	45	2.7	1	-1.733333
2	23	2.7	0	-1.73913
12	54	2.9	0	-1.740741
8	49	2.5	0	-1.755102
6	41	2.5	0	-1.756098
14	57	3	0	-2.578947
9	50	3.1	0	-2.58
15	58	3.3	0	-2.586207
16	58	3.3	0	-2.586207
10	53	3.4	0	-2.603774
18	61	3.4	0	-2.606557
5	39	3.1	0	-2.615385
13	56	3.2	0	-2.625
17	60	4.1	0	-3.466667
11	53	4.2	0	-3.471698

As the second Phat-GenIQ Model indicates, GenIQ offers more than one model solution. I present two GenIQ models to illuminate the modeler about the information imbued in the GenIQ Model score, by way of spotlighting two examples GenIQvar and GenIQvar2. The shining focused beam is on the question: Which Phat-GenIQ Model is *better*? Both models have equal predictive power, as all four males are in the top four rank positions. Accordingly, I assess the models based on two additional criteria, desirable properties of any model: *Compactness* and *discrimination* among model scores. I prefer the first Phat-GenIQ Model because it is more compact, and has greater discriminating scores than the second Phat-GenIQ Model *among the males*. The first Phat-GenIQ Model is compact, as it only has two functions (sine and subtraction). Its scores for the four males IDs #3, #4, #1 and #7 have unique discriminating GenIQvar values, +0.25638, -0.74362, -1.74622, and -1.849096, respectively. In contrast, the second Phat-GenIQ Model is not compact, as it uses seven functions, some are repeated: The basic arithmetic functions (subtraction, division, multiplication), and the Integer function. Its scores are less discriminating than the first Phat-GenIQ Model, as it assigns the same GenIQvar2 value of 0.00000 for the top two males IDs #3 and #1, respectively. Of little practical value, the third and fourth male GenIQvar2 scores are discriminating, with values of -0.888889 and -1.733333, respectively.

The less discriminating second Phat-GenIQ Model among the males readies another bright beam on whether the second model is also less discriminating (precise) than the first Phat-GenIQ Model

among the females. This inquiry can be addressed by using the coefficient of variation (CV), a measure that indicates variation, or precision in this analysis of model scores for females. The smaller the CV value means the less precise the model. Among the females, the CVs are 22.97 and 23.08 for the GenIQvar2 and GenIQvar scores, respectively. The difference in CV values is a minikin 0.11. The second Phat-GenIQ Model is *barely* less precise than the first Phat-GenIQ Model among the females. Thus, I declare the two Phat-GenIQ Models have equal precision among the females.

In summary, I prefer the first Phat-GenIQ Model over the second Phat-GenIQ Model because it is more compact, easier-on-the eyes, and has more precision among the males. Because both models have equal precision among the females, this finding is a non-issue in assessing model preference. A similar non-issue: I cannot make a comparison regarding precision between the two Phat-GenIQ Models for all subjects in the study, because CV cannot be calculated for males and females *combined* for the first Phat-GenIQ Model, yet CV can be obtained for the second model for all subjects (discussed below).

I provide a brief discussion on CV as a measure of precision of model scores. CV is only meaningful for variables with all positive/non-negative values. If *all values* under consideration are negative, then taking the absolute values of the negative values yields all positive values. I use the absolute values of both Phat-GenIQ Model scores for only females to calculate the CVs presented, above. I can calculate CV for the second Phat-GenIQ Model for *only males*; but, it serves no purpose: I cannot calculate CV for the first Phat-GenIQ Model for only males because male ID #3 has score +0.2563, and the other males have negative scores. Accordingly, I cannot calculate CV for males and females combined for the first Phat-GenIQ Model. Therefore, I cannot make a comparison based on the precision of the two Phat-GenIQ Models for all subjects in the study.

As a counterpoint to (non-physical science) analysis and modeling performed above, consider:

The world's most famous equation:
E = mc^2
It is unique, precise, and beautifully compact.

X. To Everything There is a Purpose: LRM/OLS and the GenIQ Model

The GenIQ Model, which was developed for *big data* as it lets the data define the model, is an all-powerful alternative to the statistical paradigm, requiring fitting the data to a pre-specified model. The statistical paradigm has its roots when there were only *small data*. Obviously, it still is reasonable to fit small data in a parametric pre-specified model. However, today's streaming big data in cyberspace and elsewhere require a paradigm shift. GenIQ is a utile approach for modeling big data, as big data can be difficult to *fit into* a pre-specified model. Thus, GenIQ is the *model-in-waiting* when the data – big or small – cannot fit into the regnant statistical approach. As demonstrated with the Phat Example data, GenIQ works well within small data settings.

XI. How are AGE and PERCENT_FAT related?

Some words of note about the data before I explore the relationship between AGE and PERCENT_FAT. There is a lack of consistency within the Phat Example data, as eight pairs of subjects have inconsistent data elements. Therefore, a perfect or close-to-perfect relationship between AGE and PERCENT_FAT is not possible, at least in the traditional statistics setting. However, I show that in the genetic setting a close-to-perfect relationship is obtained. This furthers the notion of GenIQs predictive prowess.

Regarding the eight pairs of subjects referenced above: The first seven subject pairs, below, have *inconsistent* data values. The last subject pair has data *consistency*, i.e., identical data values, that can add somewhat more influence on the relationship. To exhibit to the sight and mind of the reader I re-display Table 1, the Phat Example data, with highlight colors corresponding to the inconsistent/consistent data elements.

1. Subjects ID #1 and #2 have the same AGE, but different PERCENT_FAT; .
2. Subjects ID #2 and #7 have the different AGE, but same PERCENT_FAT.
3. Subjects ID #3 and #4 have the same AGE, but different PERCENT_FAT.
4. Subjects ID #5 and #9 have the different AGE, but same PERCENT_FAT.
5. Subjects ID #6 and #8 have the same AGE, but different PERCENT_FAT.
6. Subjects ID #10 and #18 have the different AGE, but same PERCENT_FAT.
7. Subjects ID #10 and #11 have the same AGE, but different PERCENT_FAT.
8. Subjects ID #15 and #16 have the same AGE and PERCENT_FAT.

Table 1. The Phat Example Data

ID	AGE	PERCENT_FAT	MALE
1	23	0.9	1
2	23	2.7	0
3	27	0.7	1
4	27	1.7	1
5	39	3.1	0
6	41	2.5	0
7	45	2.7	1
8	49	2.5	0
9	50	3.1	0
10	53	3.4	0
11	53	4.2	0
12	54	2.9	0
13	56	3.2	0
14	57	3	0
15	58	3.3	0
16	58	3.3	0
17	60	4.1	0
18	61	3.4	0

To ascertain the relationship between AGE and PERCENT_FAT I generate a plot of AGE and PERCENT_FAT in Figure 5, below. At first blush, I make the statistical evaluation that there is an underlying straight-line (linear) AGE-PERCENT_FAT relationship indicated by the red line nestled into a *tolerable* amount of random scatter. The correlation coefficient for (AGE, PERCENT_FAT) is 0.798, which implies that the quantification of the AGE-PERCENT_FAT relationship is *moderately* linear. Qualitatively, the 0.798 reflects that the observed scatter is within *reasonable* limits about the red line.

At second thought, I reset my focus on the observed scatter in the plot in Figure 5, and think that the scatter is perhaps *not* due to randomness (sample variation), but is an indicator of a source of *complexity* within the relationship. Thus, I seek out if such complexity exists. And, if so, how is it defined. I run GenIQ with PERCENT_FAT against AGE that generates the plot in Figure 6, which clearly indicates a more definite linear relationship indicated by the red line than the linear relationship in plot 5: There is *less scatter* (about the red line in Figure 6) than the scatter (about the red line in Figure 5) in the first plot. Correspondingly, I conclude that the newly genetically generated linear relationship is *accounting for* the sought-after complexity. The complexity is defined by the PERCENT_FAT-GenIQ Model. The PERCENT_FAT-GenIQ Tree and Computer Code are in Figures 7 and 8, below, following the plots. The GenIQ-based plot of AGE and

GenIQvar(PERCENT_FAT) (the latter variable is denoted by GenIQvar(PCT_FAT) in the plot) – a genetic re-expression of PERCENT_FAT – is an expected, not easy to interpret plot, which is commensurate with the ungainly interpretable GenIQ output. GenIQvar(PERCENT_FAT) as a variable in and of itself, and as one of the variables in the *genetic* plot is likely to render GP newcomers uneasy with a hard to understand variable (discussed in section XII). The correlation coefficient for (AGE, GenIQvar(PERCENT_FAT)) is 0.962, which implies that the quantification of the AGE-GenIQvar(PERCENT_FAT) relationship is *mightily* linear. Qualitatively, the 0.962 reflects that the observed scatter is within *marginal* limits about the red line.

For GP newcomers, who have receptiveness to new and different ideas, the *offspring* variable GenIQvar(PERCENT_FAT) is a *better* variable than its parent PERCENT_FAT, in that the offspring produces a stronger linear relationship in Figure 6, than PERCENT_FAT produces in the linear relationship in Figure 5. Thus, the GP newcomers' answer to the study objective about the relationship between AGE and (genetically data-mined) PERCENT_FAT is: The relationship is mightily linear with an ungainly interpretation (discussed below in section XII.) Among the GP newcomers there are *latecomers*, who cannot accept something that is not readily explicable. Their answer to the objective is: The relationship is moderately linear with clear interpretation as per the results drawn from plot in Figure 5 (discussed below in section XII).

To digress for the purpose of hurrying the latecomers to *get on board the GenIQ Express*, I relate an *expert's comment*, which at first took me aback when I heard it, but then at second thought, I was not surprised. The following question was posed on *NOVA Science Programming On-Air and Online* in 2005: How would 10 top physicists – two Nobel Prize winners among them – describe Einstein's equation $E = mc^2$ to curious non-physicists? Each physicist provided a colorful and unique interpretation (this surprised me), such that the curious non-physicist had an eclectic bouquet of ten explanations of $E = mc^2$. One physicist stated (the expert's comment): "Most people who say they understand Einstein's equation really do not understand it. Yet, they accept it." [13] Regardless of whom the physicist was referencing by *most people*, whether colleagues, the physics community at large, or the curious non-physicists, my takeaway from the expert's comment is that most people do accept *universe-al* or *true yet to-good-to-be-true* facts and findings that they may not understand, whether fully or partially. So, I hope the latecomers put *more weight* on their left-brains, just enough to let in some of the *explicable* GenIQ output in Figures 5 and 6, discussed in section XII below.
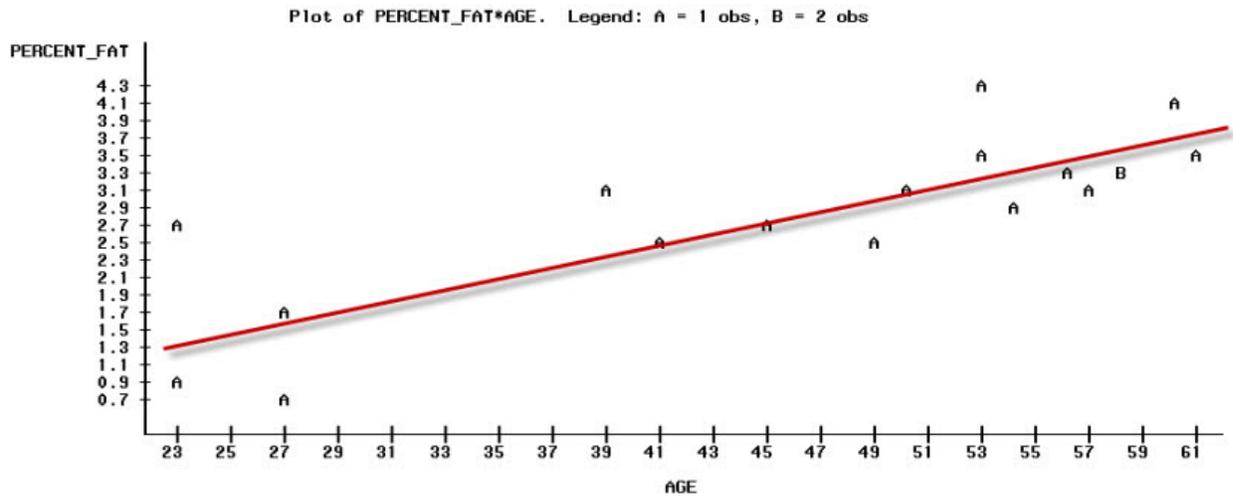
Figure 5. Plot of AGE and PERCENT_FAT

Plot of PERCENT_FAT*AGE. Legend: A = 1 obs, B = 2 obs

Figure 6. Plot of GenIQvarPCT_FAT and AGE

Plot of GenIQvarPCT_FAT*AGE. Legend: A = 1 obs, B = 2 obs, etc.
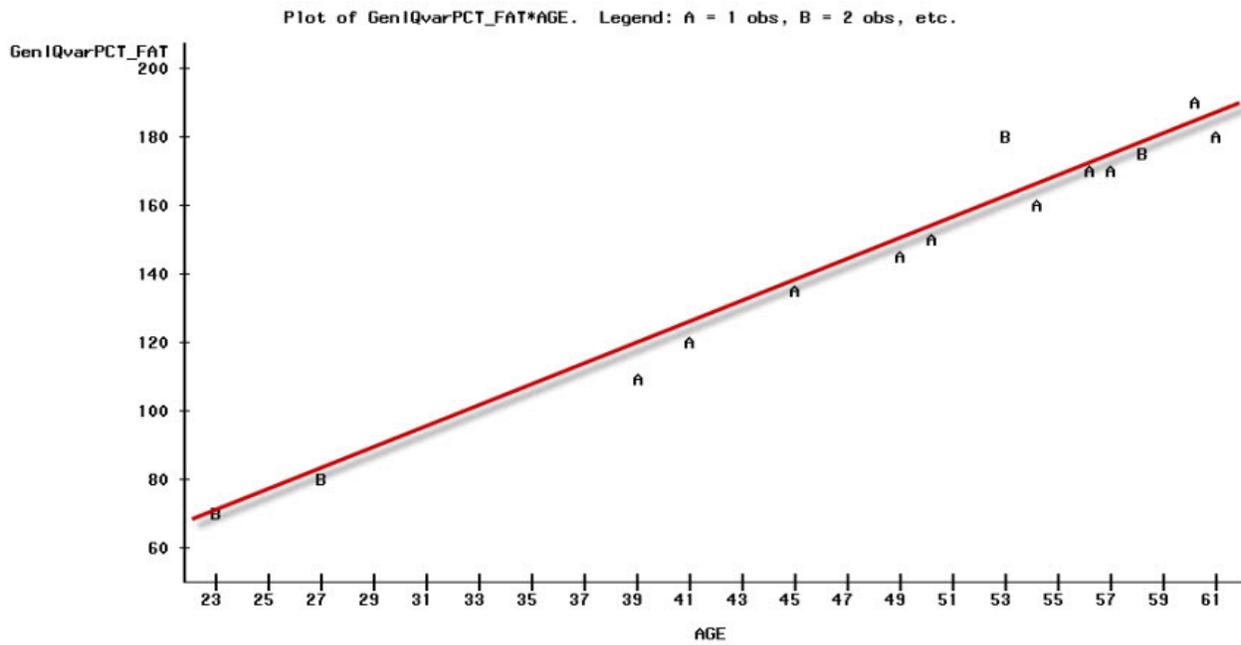
Figure 7. The GenIQ Model of PERCENT_FAT and AGE (Tree Display)
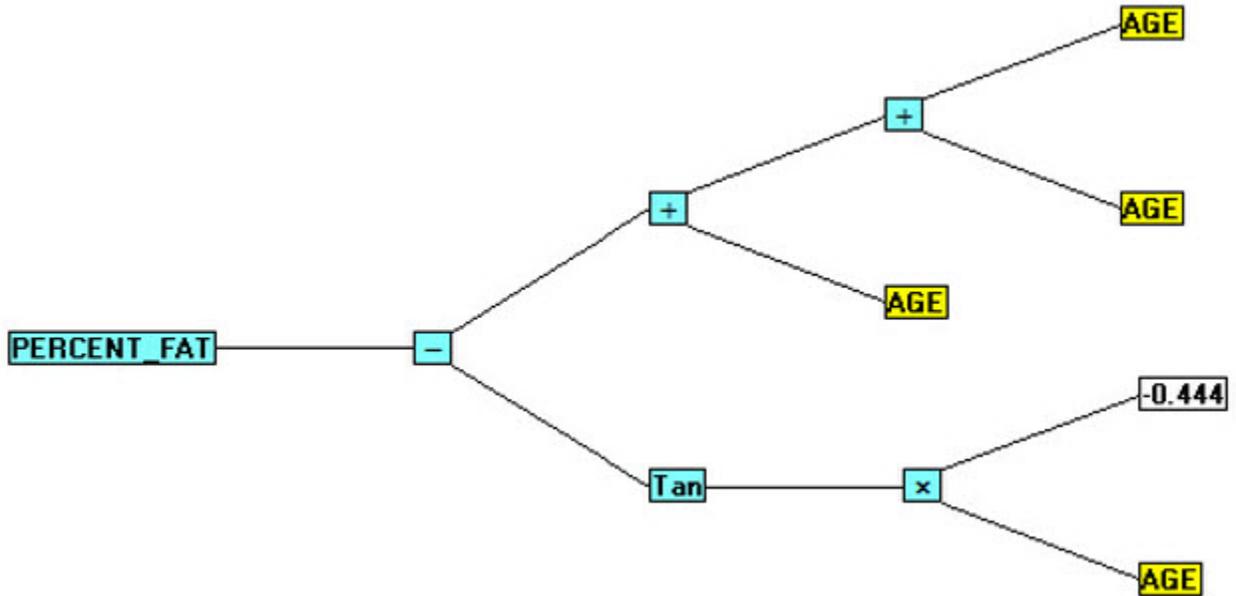

Figure 8. The GenIQ Model of PERCENT_FAT and AGE (Computer Program)

```
x1 = AGE;
    x2 = -.4436901;
  x1 = x1 * x2;
  x1 = Tan(x1);
    x2 = AGE;
        x3 = AGE;
            x4 = AGE;
        x3 = x3 + x4;
    x2 = x2 + x3;
  x1 = x2 - x1;
GenIQvar(PERCENT_FAT) = x1;
```

XII. <u>Summary of the Relationship between AGE and PERCENT_FAT</u>

The relationship between AGE and PERCENT_FAT is linear. Period. But, which version is *better*, the one based on the usual statistics assessment or the *unusual* genetic reassessment?

The usual statistics statement from PERCENT_FAT regressed on AGE is: "As AGE increases in one-year increments, PERCENT_FAT increases by a *fixed amount*, independent of the value of AGE. I perform a simple regression of PERCENT_FAT on AGE to obtain the fixed amount of 0.056 (the regression coefficient of AGE). Recall, the correlation coefficient for (AGE, PERCENT_FAT) is 0.798.

As well, I declare that the relationship between AGE and GenIQvar(PERCENT_FAT) is linear, but in an unusual genetic way. There is an obvious linear affect of AGE on GenIQvar(PERCENT_FAT) in the plot in Figure 6 that requires attention to describe. To paraphrase Shakespeare, "What's in a linear relationship? That which we call a linear relationship by any other name would appear to be a linear relationship." Consequently, the Shakespearean-minded modeler declares that the AGE-GenIQvar(PERCENT_FAT) relationship is *genetically linear,* accounting for the complexity that is attributed to the scatter in the plot in Figure 5. Thus, the unusual genetic statement is: "As AGE increases in one-year increments, *GenIQvar(PERCENT_FAT) increases* by a fixed amount, independent of the value of AGE. I perform a simple regression of GenIQvar(PERCENT_FAT) on AGE to obtain the fixed amount of 3.231 (the regression coefficient of AGE). Recall, the correlation coefficient for (AGE, GenIQvar(PERCENT_FAT)) is a mighty 0.962. So, *which version is better?* I declare the AGE-GenIQvar(PERCENT_FAT) relationship the better version because its correlation coefficient is significantly larger than that of the first AGE-PERCENT_FAT relationship. After all, *that which we call a linear relationship by any other name would appear to be a linear relationship.*

Not to overlook the unusual genetic result of GenIQvar(PERCENT_FAT) increases by the fixed amount of 3.231 for any one-year increase in AGE: The metric structure of GenIQvar(PERCENT_FAT) requires an analytical study of the genetic variable, which may seem like a towering task. In fact, the opposite is true. The modeler simply selects, say, three two-year pairs of ages (e.g., 23 and 24; 34 and 35, and 55 and 56), calculates GenIQvar(PERCENT_FAT) values from the computer code (lest one forget the computer code is just an equation), and then plots the three pairs of ages (on the x-axis) and the corresponding GenIQvar(PERCENT_FAT) values (on the y-axis). The resultant plot provides a visual picture, worthy of a thousand words (from which I am spared), which explains the mighty linear affect of AGE on GenIQvar(PERCENT_FAT).

XIII. <u>The Hybrid Phat LRM-GenIQ Model</u>

At the outset of this article, I put forward a trinity of the regression paradigm, which opens up an additional alternative model to the statistical regression models. The alternative model is based on: The Hybrid Statistics-GP paradigm – integrating the best characteristics of two paradigms – yields a utile alternative of LRM/OLS or GenIQ. For this article, the hybrid paradigm is: The modeler fits the Phat Example data to a LRM with the modeler's preferred variable selection method to determine the best subset among the original and GenIQ *genetically* data-mined variables. Of

primary import, the Hybrid Phat LRM-GenIQ Model includes the regression coefficients, which provide the necessary comfort level of model interpretability for model acceptance. In view of the foregoing, I build a Hybrid LRM-GenIQ Model discussed below.

The Hybrid Phat LRM-GenIQ modeling effort considers the candidate predictor variable set of four variables: two from the preferred first Phat-GenIQ Model, sine_of_AGE and GenIQvar, and the study's two original variables, AGE and PERCENT_FAT. The target variable is MALE as previously defined in section I. The Hybrid Phat LRM-GenIQ Model uneventfully turned out identical to the already built Phat-LRM in section III. The nonoccurrence of a different model for the hybrid model is due to the data condition of multicollinearity, namely, the variables under consideration are highly correlated. (Multicollinearity is problematic for statistical regression models, but not for the GenIQ Model. [14]) For the Hybrid LRM/OLS-GenIQ Model, it is a moderate-to-severe degree of multicollinearity among the candidate predictor variable set that affects the hybrid model. Non-intuitive to the statistician, this result is *not* typical when building a Hybrid LRM/OLS-GenIQ Model with *many original* variables. [15] When there are many original variables, there are many genetically data-mined variables. This results in a large candidate predictor variable subset such that the condition of multicollinearity *lessens* to yield a high-level predictive Hybrid LRM/OLS-GenIQ Model. For the Hybrid Phat LRM-GenIQ Model, the study's small number of original variables (i.e., two) creates a greater-than-moderate degree of multicollinearity among candidate predictor variable set. This resulted in the Hybrid Phat LRM-GenIQ Model identical to the Phat-LRM.

XIV. Conclusion

I present a new genetic programming model – the GenIQ Model – an alternative model to the statistical ordinary least squares and logistic regression models. First, I demonstrate the GenIQ Model's theoretical superiority and usable advantages over the long-standing, widely used statistical regression models. GenIQ lets the data define the model, and as such it presupposes big data under consideration. GenIQs genetic data-defining paradigm is an exceedingly utile alternative to the statistical fit-the-data paradigm, especially for big data environment. Today's streaming big data in cyberspace and elsewhere require a paradigm shift away from the statistical paradigm, which has its roots when there were only small data. The statistical paradigm, long-standing and widely used, is still obviously reasonable to fit small data in the LRM/OLS parametric pre-specified model. However, GenIQ is a valuable approach with predictive prowess and usable features for modeling big data, as big data can be difficult to *fit into* a pre-specified model. Thus, GenIQ is the *model-in-waiting* when the data – big or small – cannot fit into the regnant statistical approach. As demonstrated with the Phat Example data, GenIQ works well within small data settings.

Second, I expound and illustrate the GenIQ features: 1) Data mining among the original variables, 2) Performing variable selection by deciding the best subset of predictor variables, among the original and GenIQ genetically data-mined variables, and 3) Setting forth the model equation itself – such that 4) The decile table, a measure of model performance, is optimized. LRM/OLS has no data mining capability. And, LRM/OLS variable selection methods are too numerous to mention, and viewed by many statisticians as "unclean and distasteful." The implication is that there is no *best* variable selection approach to yield a *best* LRM/OLS. [16] LRM/OLS fitness functions serve

as surrogates for explicit fitness functions for decile optimization. I highlight the GenIQ singular weakness of difficulty in interpreting GenIQ output that consists of completing two parts – the equation, an unsuspected computer program, and a visual display with Picasso-like abstractness of the program – that account for the limited use of GenIQ. As an antithetically difference from GenIQ, LRM/OLS is interpretable, as made possible by the redoubtable regression coefficients.

I use a small, yet difficult data (because of inconsistent data values for several subject-pairs) to show the extraordinary predictive ability, and versatility of the GenIQ Model. The data, which I dub the Phat Example data, come from a study investigating a new method of measuring body composition, consist of three predictor variables: Body fat percentage (PERCENT_FAT), AGE, and gender/MALE (there are four males) for eighteen normal adults aged between 23 and 61 years. The study has two objectives: Is there any evidence that the relationship is different for males and females? And, How are AGE and PERCENT_FAT related? I build two MALE classification models, which predict perfectly the likelihood of a subject's gender is a male, thus providing the evidence. For the second objective, I generate several bivariate plots to assess whether or not the observed scatter in the first plot of AGE and PERCENT_FAT is due to sample variation, or is an indictor of a source of complexity within the relationship between the two variables. I use GenIQ to confirm that the observed scatter is a source of complexity within the relationship, and consequently, to define the complexity.

Using the Phat Example data, I build a Phat-LRM, two Phat-GenIQ Models, and a Hybrid Phat LRM-GenIQ Model. The best model is the first Phat GenIQ Model: It optimizes the likelihood of a subject's gender is male. The second Phat GenIQ Model also optimizes likelihood of a subject's gender is male, but is neither compact nor precise as the first Phat-GenIQ Model.

I propose that the GenIQ modeling approach opens up an additional alternative hybrid model paradigm to LRM/OLS: The Hybrid LRM/OLS-GenIQ paradigm consists of the modeler fitting the data to LRM/OLS with the modeler's preferred variable selection method to determine the best subset among the original and GenIQ genetically data-mined variables. The Hybrid LRM/OLS-GenIQ Model includes undoubtedly highly predictive genetically data-mined variables. The Hybrid Phat LRM-GenIQ Model considers the candidate predictor variable set of two genetically data-mined variables from the preferred first Phat-GenIQ Model, and the study's two original variables. The Hybrid Phat LRM-GenIQ Model is identical to the Phat-LRM. The nonoccurrence of a different model for the hybrid model is due to the greater-than-moderate degree of multicollinearity resulting from the study's small number of original variables. Non-intuitive to the statistician, this result is *not* typical when building a Hybrid LRM/OLS-GenIQ Model with *many original* variables. When there are many original variables, there are many genetically data-mined variables. This results in a large candidate predictor variable subset such that the condition of multicollinearity *lessens* to yield quality predictive Hybrid LRM/OLS-GenIQ Models.

Regarding the second objective: How are AGE and PERCENT_FAT related? I generate the plot of PERCENT_FAT against AGE. I assess that the data suggests an underlying linear relationship between the two variables. But, I reset my focus on the observed scatter in the latter plot, and believe that the scatter is perhaps not due to randomness, but is an indicator of a source of complexity within the relationship. I run GenIQ with target variable PERCENT_FAT against AGE to uncover a mighty linear relationship, accounting for the complexity of the relationship. The

GenIQ-based plot of AGE and GenIQvar(PERCENT_FAT) does indeed capture the complexity of the relationship, as the observed relationship now suggests a more definitive underlying linear relationship between AGE and GenIQvar (PERCENT_FAT).

I present two versions of the linear relationship between AGE and PERCENT_FAT. The traditional linear relationship defined by a regressing PERCENT_FAT on AGE. The second version is a genetic linear relationship defined by regressing GenIQvar(PERCENT_FAT) on AGE. I declare the AGE-GenIQvar(PERCENT_FAT) relationship the better version because its correlation coefficient (0.962) is significantly larger than the correlation coefficient (0.798) of the first AGE-PERCENT_FAT relationship.

In summary, the GenIQ Model is the appropriate model where the decile table is the unquestionable measure of model performance. For other instances, a trade-off has to be made between GenIQs predictive ability, usable features, and no coefficients (i.e., the output that provides model interpretability) versus the statistical LRM/OLS interpretability, and their fitness functions as surrogates for optimizing the decile table. Separation anxiety from something that it long-standing, and widely used for centuries is a condition that takes time to treat. Until the checks (i.e., the equation is a computer program and a visual Picasso-like display of the program) become ocularly palatable, which comes about with retraining statisticians to think *out-of-the-box*, LRM/OLS will continue to be used (in big data settings). With the eventual recognition that *valuable information* comes in unsuspected forms (the GenIQ Tree Display), GenIQ will be popular.

GenIQ is the model for today's data. It can accommodate big and small data, as it is an assumption-free, nonparametric flexible (i.e., no pre-specification) of the model, whose genetic programming paradigm lets the data define the model. In stark contrast, LRM/OLS were conceived, testing and experimented on the small data setting of their day. These models are suboptimal and problematic with today's big data. [16, 17] The LRM/OLS paradigm is to *fit the data to* an assumption-full, parametric inflexible (i.e., pre-specified) model, which is best for small data settings. GenIQ will become popular as today's data grows in size, necessitating the data define the model, rather than *fitting square data into a round model.*

References
1. GenIQ Model© (www.GenIQModel.com)
2. GP paradigm (http://www.geniq.net/Koza_GPs.html)
3. Historical Notes on the Two Most Popular Prediction Models, and One Not-yet Popular Model (http://www.geniq.net/res/HistoricalNotesTwoMostPopularModelsOneNotYet.html)
4. The GenIQ Model was conceived, tested and experimented with big data. As a model that is data-defined, GenIQ presupposes the data under consideration are big data.
5. Ratner, B., *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, CRC Press, Boca Raton, 2003, Chapters 15, 16 and 17.
6. Decile Analysis Primer (http://www.geniq.net/res/DecileAnalysisPrimer_2.html)
7. The "Smart" Decile Analysis (http://www.geniq.net/res/SmartDecileAnalysis_.html)
8. Harlow, L. L., Mulaik, S. A., Steiger, J. H., *What If there Were No Significance Tests?,* Lawrence Erlbaum Associates, Publishers, Mahwah, New York, 1997.
9. Miller, A., J., *Subset Selection in Regression*, Chapman and Hall, London, 1990.

10 *Extra-GenIQ Applications* (http://www.geniq.net/Extra-GenIQ-Applications.html)

11. American Journal of Clinical Nutrition, 40, 834-839.

12. Ratner, B., *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, CRC Press, Boca Raton, 2003, Chapter 6.

13. NOVA Science Programming On-Air and Online, *Einstein's Big Idea: E=mc^2 Explained* (http://www.pbs.org/wgbh/nova/einstein/experts.html), 2005.

14. Statistical Modeling Problems: Nonissue for GenIQ (http://www.geniq.net/res/statistical-modeling-problems-nonissue-for-GenIQ.html)

15. Ratner, B., *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, CRC Press, Boca Raton, 2003, Chapters 3 and 4.

16. Miller, A., J., *Subset Selection in Regression*, Chapman and Hall, London, 1990.

17. Ratner, B., *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, CRC Press, Boca Raton, 2003, Chapters 3 and 4.