GenIQ: A Method that Lets the Data Specify the Model

Bruce Ratner PhD

DM STAT-1 CONSULTING

1 800 DM STAT-1 www.DMSTAT1.com

GenIQ©

Typically, the data analyst approaches a problem directly with an "inflexible" designed procedure specifically for that purpose.
For example, the statistical problem of predicting a continuous target variable (e.g., profit) is solved by the "old" classical standard ordinary least-squares (OLS) regression model.

Atypically, the new machine learning GenIQ Model©, a "flexible" nonparametric, assumption-free approach, *Lets the Data Specify the Model (equation)*.

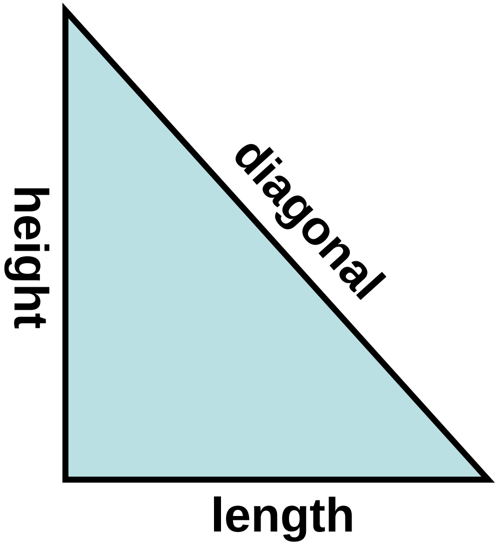
Specifically, GenIQ *let's the data specify the model* automatically –

1. **data mines for new variables among the original vars.**
2. **performs variable selection among the new & original**
3. **specifies the model equation.**

OBJECTIVE: To show how GenIQ *"Lets the Data Specify the Model."*

Consider the well-known the Pythagorean Theorem. Given a right triangle with the three sides denoted by

l = length
h = height
d = diagonal (hypotenuse)



Pythagorean Theorem:

$$d^2 = l^2 + h^2, \text{ equivalently}$$

$$d = \sqrt{(l^2 + h^2)}$$

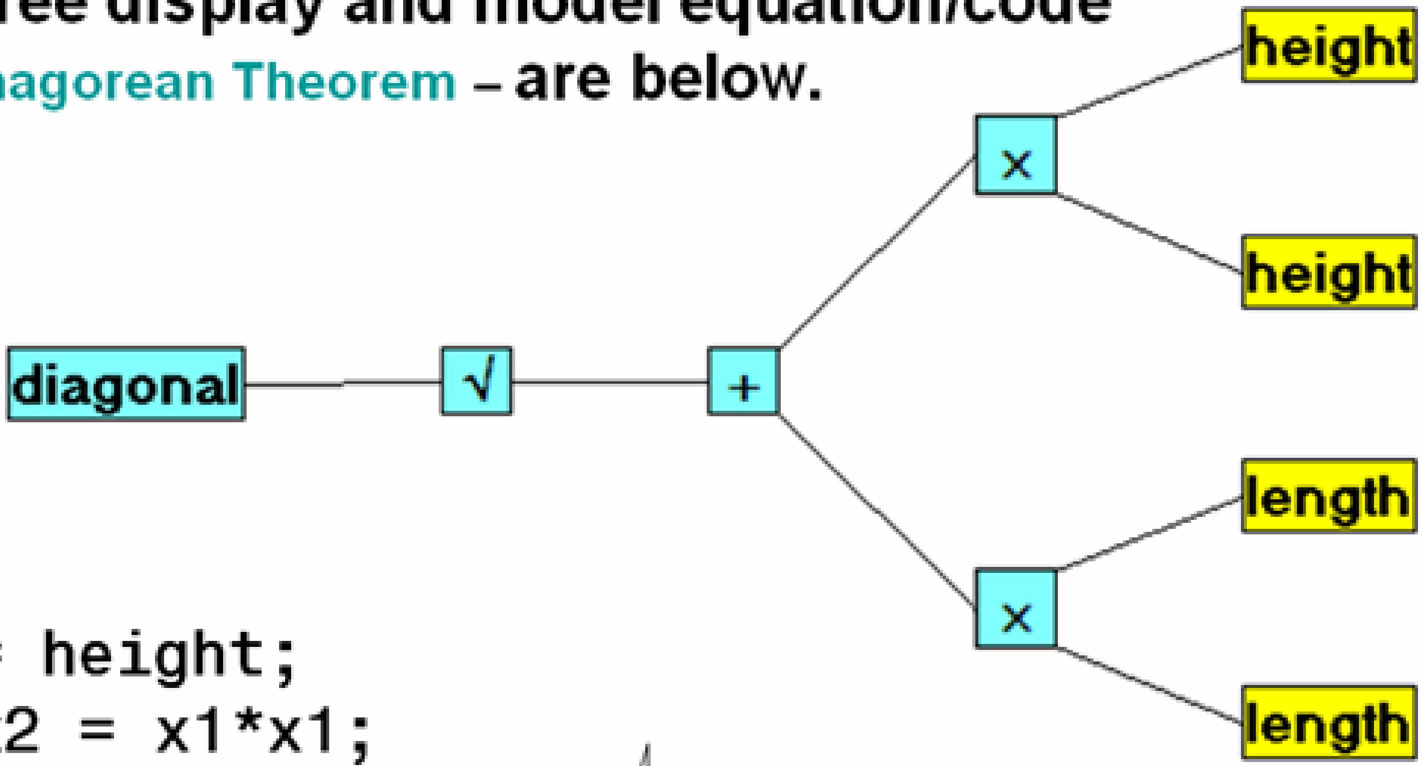
INPUT: GenIQ requires selection of variables and functions.
Variables: Ten pythagorean triplets (l,h,d) in Table 1, below.
Functions: Arithmetic (+, -, %, x), and the SQRT (square root).

Table 1

(l,h,d)	length	height	diagonal
1	3	4	5
2	5	12	13
3	7	24	25
4	8	15	17
5	9	40	41
6	11	60	61
7	12	35	37
8	13	84	85
9	16	63	65
10	20	21	29

RESULTS:

GenIQ finds the relationship among the pythagorean triplets. GenIQs tree display and model equation/code – the Pythagorean Theorem – are below.



```
x1 = height;
   x2 = x1*x1;
      x3 = length;
         x4 = x3*x3;
   x5 = x2 + x3;
x6 = SQRT(x5);
diagonal = x6;
```

The model code as displayed is standard. But, assuredly, it is the Pythagorean Theorem defining the diagonal.

Notwithstanding how GenIQ finds the pythagorean formula, one can argue that GenIQ is not put to the task. In this case, because the answer is known, SQRT is included with the arithmetic functions.

The Contention:

Can GenIQ produce a robust, accurate, and stable model via *Letting the Data Specify the Model* when the correct functions are not selected?

OBJECTIVE #2:

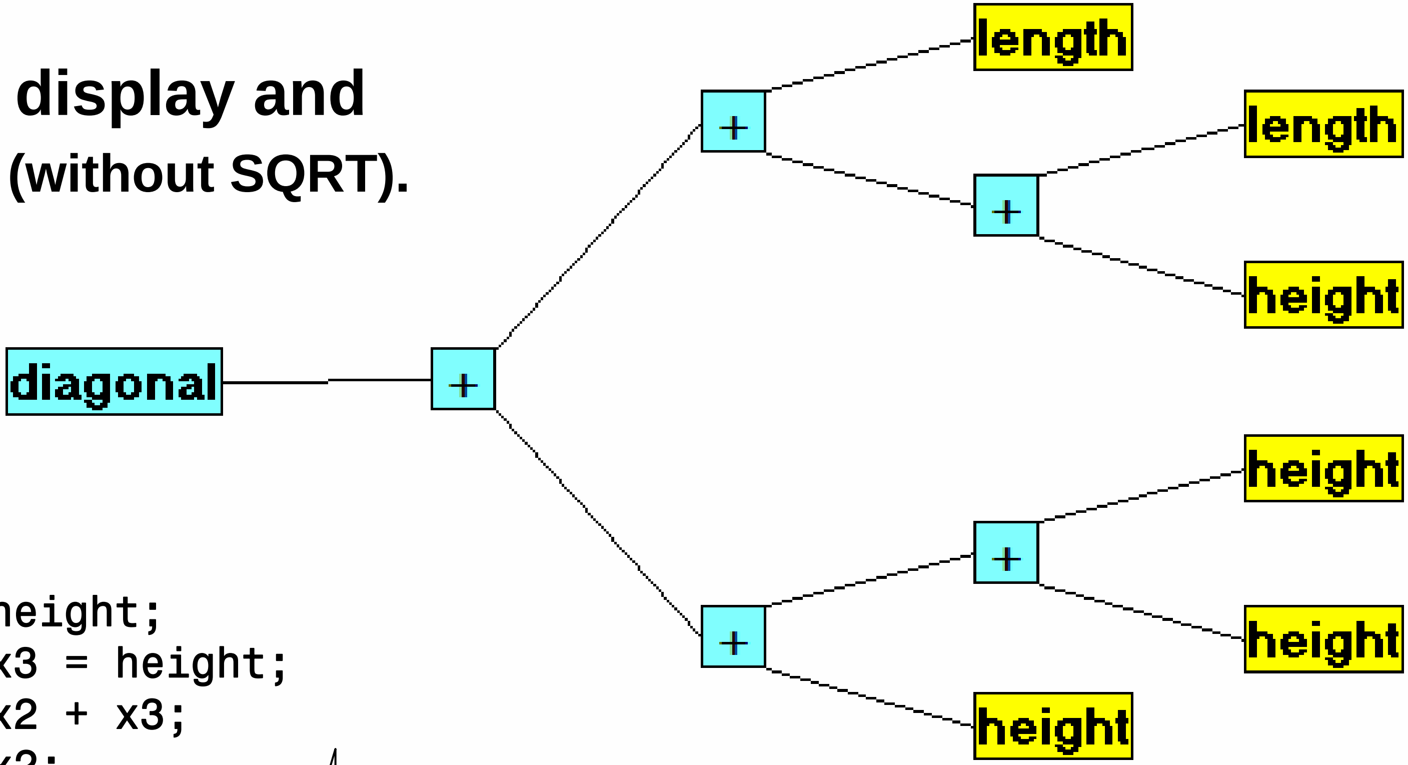
How does GenIQ fair when the correct INPUT is not known, viz., when only arithmetic functions are selected?

INPUT: I continue with the Pythagorean illustration.

Variables: Ten pythagorean triplets (l,h,d) in Table 1.

Functions: Only arithmetic (+, -, %, x).

RESULTS:
GenIQs tree display and model code (without SQRT).



```
 x1 = height;
          x2 = height;
                x3 = height;
          x2 = x2 + x3;
     x1 = x1 + x2;
          x2 = height;
                x3 = length;
          x2 = x2 + x3;
                x3 = length;
          x2 = x2 + x3;
     x1 = x1 + x2;
GenIQvar = x1;
GenIQ_diagonal_Score = -0.6296496 + 0.2343271 * GenIQvar;
```

The model code as displayed is standard.
In this case, it is an alternative Pythagorean Theorem defining the diagonal as a GenIQ estimate:
diagonal = 4*height + 2*length

➢ GenIQ without SQRT performs kinda nicely, discussed below.

✓ The implication: GenIQ is not burdened by not knowing the unknowable (correct) functions.

➢ GenIQs predictive power and data mining prowess produce - via its evolutionary process of genetic programming - new variables (along with the other tree branches) that contribute to the predictive accuracy of the otherwise unknown “correct” functions.

➢ I calculate the following in Table 2, below:

1. ERROR = diagonal - GenIQ_diagonal_Score
2. Abs_ERROR = absolute (diagonal - GenIQ_diagonal_Score)
3. Squared Error (SE) = (diagonal - GenIQ_diagonal_Score)**2
4. Absolute % Error (APE) = (Abs_ERROR / diagonal)*100

- **I observe modest means of Abs_ERROR, SE and APE:**

```
1. Mean APE (MAPE) = 0.0185
2. Mean SE (MSE)   = 0.2800
3. Mean Abs_ERROR  = 0.4251
```

- **It is concluded that *in toto* GenIQ produces a rather robust, accurate *, and stable** model as it:**

Lets the data specify a close-to-true model.

* - See Table 2, page 10.

** - I concede that this illustration has not addressed stability.

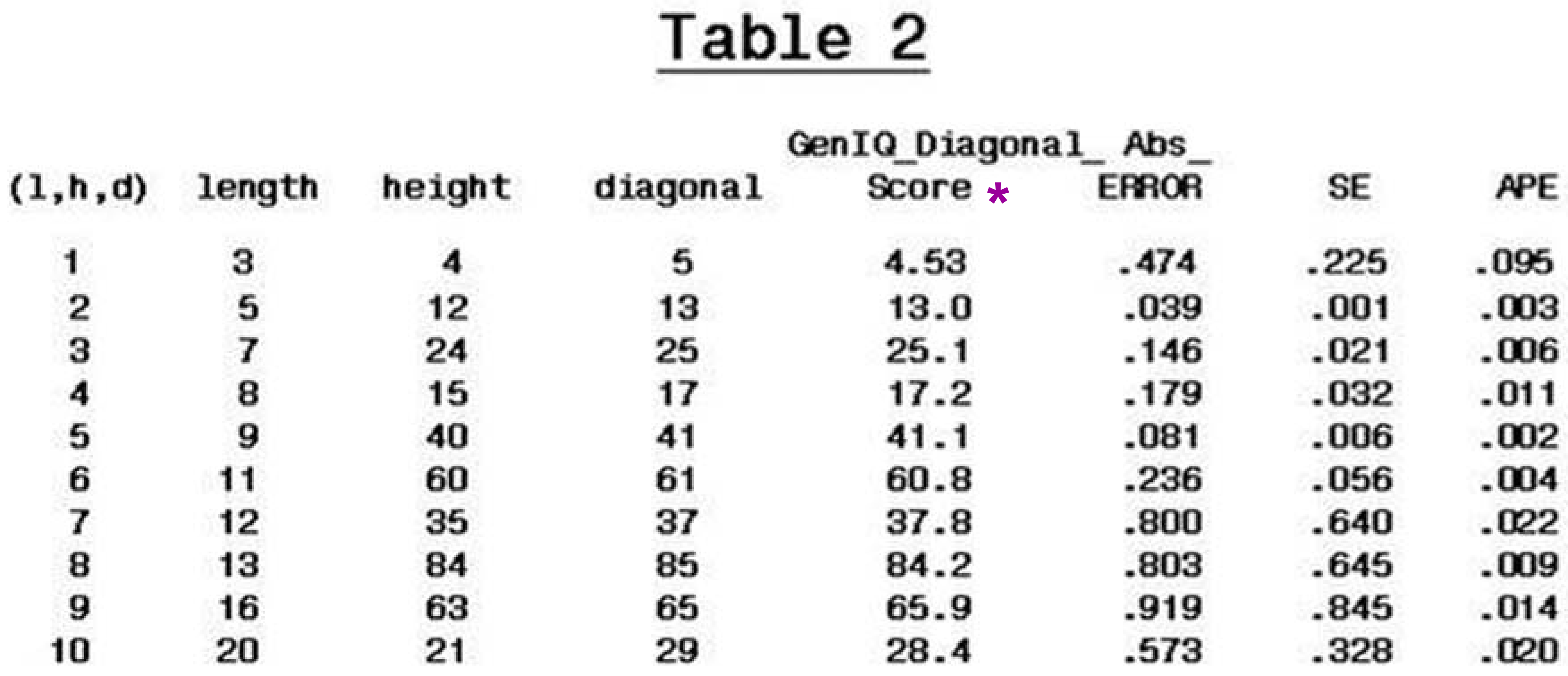


Table 2

(l,h,d)	length	height	diagonal	GenIQ_Diagonal_ Score *	Abs_ ERROR	SE	APE
1	3	4	5	4.53	.474	.225	.095
2	5	12	13	13.0	.039	.001	.003
3	7	24	25	25.1	.146	.021	.006
4	8	15	17	17.2	.179	.032	.011
5	9	40	41	41.1	.081	.006	.002
6	11	60	61	60.8	.236	.056	.004
7	12	35	37	37.8	.800	.640	.022
8	13	84	85	84.2	.803	.645	.009
9	16	63	65	65.9	.919	.845	.014
10	20	21	29	28.4	.573	.328	.020

I would greatly appreciate your comments about GenIQ Letting the Data Specify the Model. Please email me.

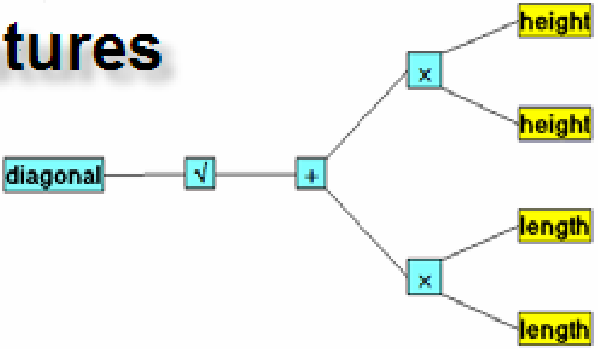
Thank you. Bruce

* **= GenIQ_Diagonal_Score is the predicted diagonal value from the second GenIQ Model**

Preview of GenIQ Features

To show that GenIQ:

1. Lets the data define the model
2. Data mines for new variables
3. Performs variable selection, and
4. Specifies the model equation –
5. So as to optimize the *decile table* .



- The GenIQ output for the Pythagorean Theorem is not so ungraspable because the solution is well known; it is the sixth most famous equation.

- GenIQs tree and computer code represent **features #1 and #4:**
  - GenIQ Model is a machine learning (ML) regression method that specifies the model by letting the data define the model.

- GenIQ tree represents **feature #2** of GenIQ:
  - The GenIQ Model automatically data mines for new variables.
  - There are four new variables (branches):
    New_var1=(height X height)
    New_var2=(length X length)
    New_var3=(New_var1 + New_var2), and lastly,
    New_var4=SQRT(New_var3), which is the model itself.

Preview of GenIQ Features

- Thus, the GenIQ Model serves as a unique data mining method creating new variables – that can not be intuit – via the GP process, which evolves structure (new vars.)
  - "without explicit programming."
- Moreover, appending the new variables to the dataset with the original variables for building a statistical regression model,
  - Produces a hybrid statistics-ML model,
  - Along with the regression coefficients that provide the regression modeler the necessary comfort level for model acceptance.