Building Statistical Regression Models: Straight Data are Necessary

~ GenIQ as a Data Straightener ~

Bruce Ratner PhD

DM STAT-1 CONSULTING

1 800 DM STAT-1 www.DMSTAT1.com

GenIQ©

~ The GenIQ Model ~

The GenIQ Model© is a machine learning alternative model to the statistical ordinary least squares and logistic regression models. GenIQ lets the data define the model – automatically data mines for new variables, performs variable selection, and then specifies the model equation – so as to "optimize the decile table," to fill the upper deciles with as much profit/many responses as possible.

In this illustration, GenIQs optimizing of the deciles is equivalent to maximizing the correlation between the target variable and the GenIQs straighten predictor variable, **GenIQvar**.

The quotidian techniques for predicting continuous and binary target variables are the statistical ordinary least squares (OLS), and logistic regression models, respectively. These regression methods are called *linear* models because the target variable (Y) is expressed as a "linear combination" of the regression coefficients (b_i), i.e., as a weighted sum of the predictor variables (X_1, X_2, …, X_n):

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

- where the weights are the regression coefficients b_i, and X_0=1.

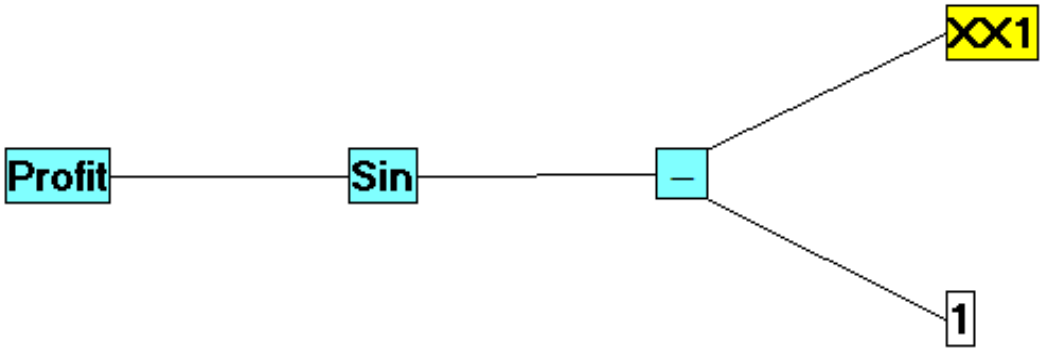
The linearity assumption implies a necessary condition for building statistical regression models: Each predictor variable has a linear or straight-line relationship with the target variable. A popular technique for straightening data is Tukey's Bulging Rule (TBR). However, if the predictor-target variable relationship has a "kink," the TBR will not work. Notwithstanding kinks, TBR must be applied to each and every predictor variable, one at a time against the target variable. TBR is very time-consuming, and limited to data with a bulge, not a kink.

The purpose of this illustration is to present the ***GenIQ Model as a data-straightener***, which is robust, powerful, and without restrictions or limitations. Moreover, GenIQ can be applied to many variables – at one time. GenIQ is automatic, and thusly a time-saver.

OBJECTIVE: To build a **GenIQ Model to predict Profit** using XX1 with data in Table 1. The GenIQ Model tree display and model equation (code) are below.

Table 1

ID	XX1	Profit
1	-1.0	1.00
2	-0.8	0.84
3	-0.6	0.76
4	-0.4	0.76
5	-0.2	0.84
6	0.0	1.00
7	0.2	1.24
8	0.4	1.56
9	0.6	1.96
10	0.8	2.44
11	1.0	3.00




```
x1 =  1;
   x2 = XX1;
      x1 = x2 - x1;
   x1 = Sin(x1);
GenIQvar = x1;
```

GenIQ RESULTS: The Profit ranking, based on **GenIQvar** in Table 3 below, **is perfect**! Label letters go from **a** "straight" to **k**. The Label letters assist in identifying the perfect-order of Profit: **a**, **b**, …, **k**. (GenIQvar is a ***unitless number:*** the larger the value the greater the contribution of profit.)

Table 3

Label	XX1	Profit	GenIQvar
a	1.0	3.00	0.00000
b	0.8	2.44	-0.19867
c	0.6	1.96	-0.38942
d	0.4	1.56	-0.56464
e	0.2	1.24	-0.71736
f	0.0	1.00	-0.84147
g	-1.0	1.00	-0.90930
h	-0.2	0.84	-0.93204
i	-0.8	0.84	-0.97385
j	-0.4	0.76	-0.98545
k	-0.6	0.76	-0.99957

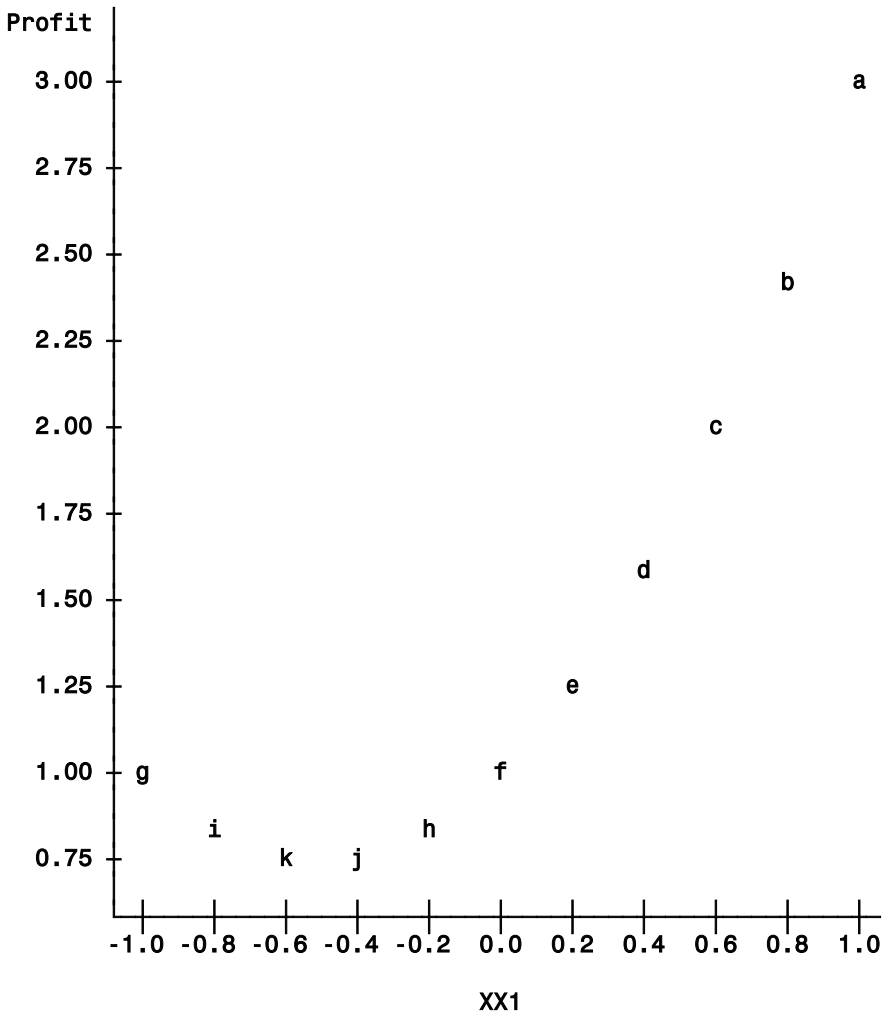
As previously mentioned, GenIQs optimizing of the deciles is equivalent to maximizing the correlation between the target variable and the GenIQs straighten predictor variable, **GenIQvar**. Thus, the best way of illustrating GenIQ as a data-straightener is to assess the visual and numeric displays in the following slide #9:

- the plots of: Profit versus XX1, and **GenIQvar** versus XX1, and
- the correlations between: Profit with XX1, and **GenIQvar** with XX1.

Plot Profit versus XX1 clearly shows a non straight-line relationship between Profit and XX1. Thus, the corresponding correlation coefficient of 0.87304 is not a valid measure. Interesting to note: The shape of the relationship is not totally "smooth." This can be seen as one connects the Profit-letters in alphabetical order: From **a** to **f** is smooth. However, there is a zigzag pattern starting at **f**: from **f** to **g** (left) to **h** (down right) to **i** (left) to **j** (down right) to **k** (left).

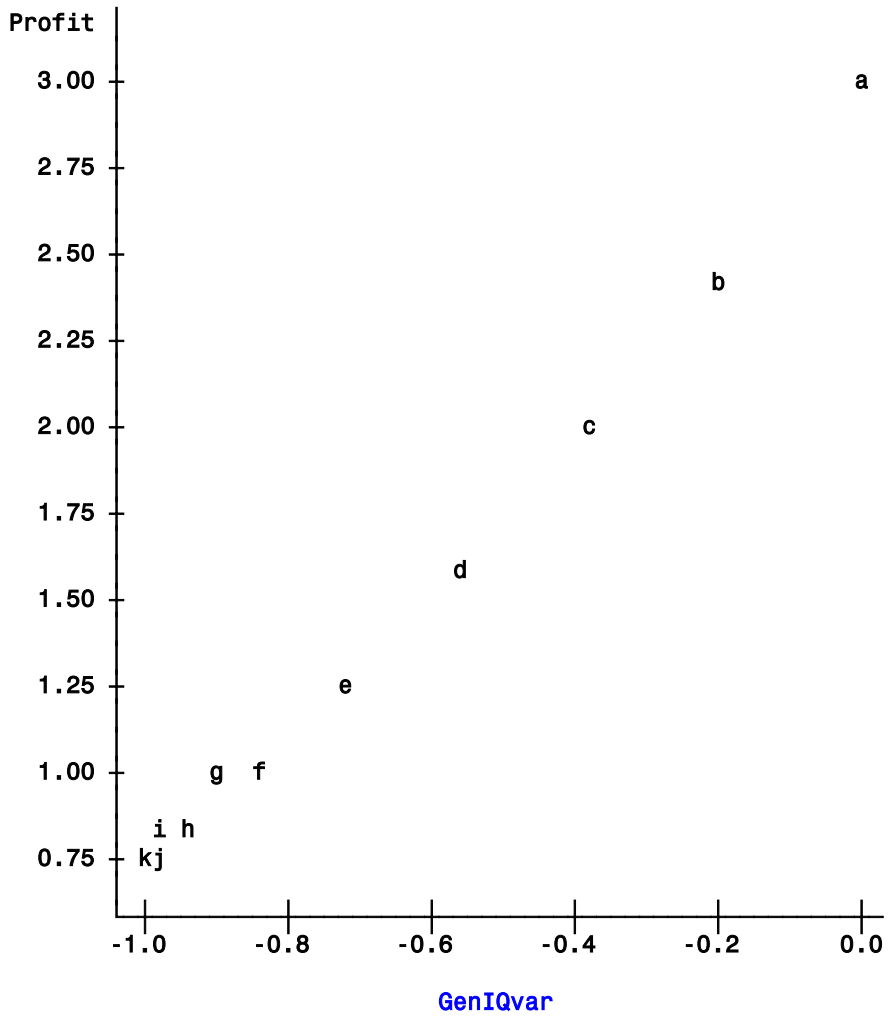
Plot Profit versus **GenIQvar** clearly shows a straight-line relationship between Profit and **GenIQvar**. Thus, the corresponding correlation coefficient of 0.99514 is valid, and indicates a very strong straight-line relationship, i.e., the data are straight! Interesting to note: There are three couplets of equal Profit values - (**f** and **g**, Profit = 1.00), (**h** and **i**, Profit = 0.84), and (**j** and **k**, Profit = 0.76) - for which GenIQ discriminates nicely by assigning diverse GenIQvar scores ranging from -0.84147 to -0.99957.

Plot of Profit*XX1.



Plot of Profit*GenIQvar.

Symbol is value of label for perfect-order Profit: a,..., k.



Correlation Coefficients

	XX1	GenIQvar
Profit	0.87304	0.99514

I would greatly appreciate your comments about GenIQ as a data-straightener.
Please email me.
Thank you. Bruce