
A Method for Moderating Outliers ~ Instead of Discarding Them ~

Bruce Ratner PhD
DM STAT-1 CONSULTING
| 800 DM STAT-1 | www.DMSTAT1.com

GenIQ[®]

Objectives

- To present a method - the **GenIQ Model**© - for moderating outliers, instead of discarding them.
 - **GenIQ** transforms the independent variable(s) to straighten a given relationship, effectively moderating the outlier(s).
- To illustrate **GenIQ** as a method for handling outliers with a simple dataset.
- **GenIQ** is especially useful for building ordinary least squares and logistic regression models, as these models are sensitive to outliers.

Outliers: What They Are? What To Do?

- In statistics an outlier is an observation that lies outside the overall pattern of the rest of the data.
- Outliers can also occur when comparing relationships between two or more variables.
 - Outliers of this type are easily identified on a scatterplot.

Outliers: What They Are? What To Do?

- Disgarding outliers is a controversial practice frowned on by many statisticians and data analysts.
 - While mathematical criteria provide an objective and quantitative method for data rejection, they do not make the practice more scientifically or methodologically sound.

Perfect Correlation Coefficient ... If Not for a Single Outlier

- Consider the dataset of 101 points (XX, Y) .
 - There are four "mass" points, each have 25 observations:
 - ▶ $(17, 1)$ has 25 observations
 - ▶ $(18, 2)$ has 25 observations
 - ▶ $(19, 4)$ has 25 observations, and
 - ▶ $(20, 4)$ has 25 observations.
 - There is one "single" point.
 - ▶ $(1, 20)$ has 1 observation.

Correlation Coefficient & Scatterplot

- I calculate the correlation coefficient and generate the required scatterplot,
 - which is a necessary visual display to check the Correlation Coefficient Linearity Assumption.

Correlation Coefficient & Scatterplot (*continued*)

- Linearity Assumption
 - The correlation coefficient **requires** that the underlying relationship between the two variables under consideration is **linear**.
 - If the relationship is known to be linear, or the observed pattern between the two variables appears to be linear,
 - ▶ then the correlation coefficient provides a **reliable** measure of the strength of the linear relationship.
 - If the relationship is known to be nonlinear, or the observed pattern appears to be nonlinear, then the correlation coefficient is **not useful**, or at least questionable.

Perfect Correlation Coefficient ... If Not for a Single Outlier

- The correlation coefficient of (XX, Y) is -0.41618 , in Table 1.
- The corresponding scatterplot, in Figure 1, shows that the single *point (1, 20)* is clearly *an outlier*, and the assumption of linearity *does not hold*.
 - The relationship between XX and Y is not linear.
- Thus, the *correlation coefficient* of -0.41618 is *not reliable* measure of the relationship between XX and Y .

Perfect Correlation Coefficient ... If Not for a Single Outlier

Table 1. Correlation Coefficients

Pearson Correlation Coefficients, N = 101
Prob > |r| under H0: Rho=0

	XX	GenIQvar
Y	-0.41618 <.0001	0.84156 <.0001

Perfect Correlation Coefficient ... If Not for a Single Outlier

Plot of Y*XX. Legend: A=1 obs, B=2 obs, ..., Y=25 obs

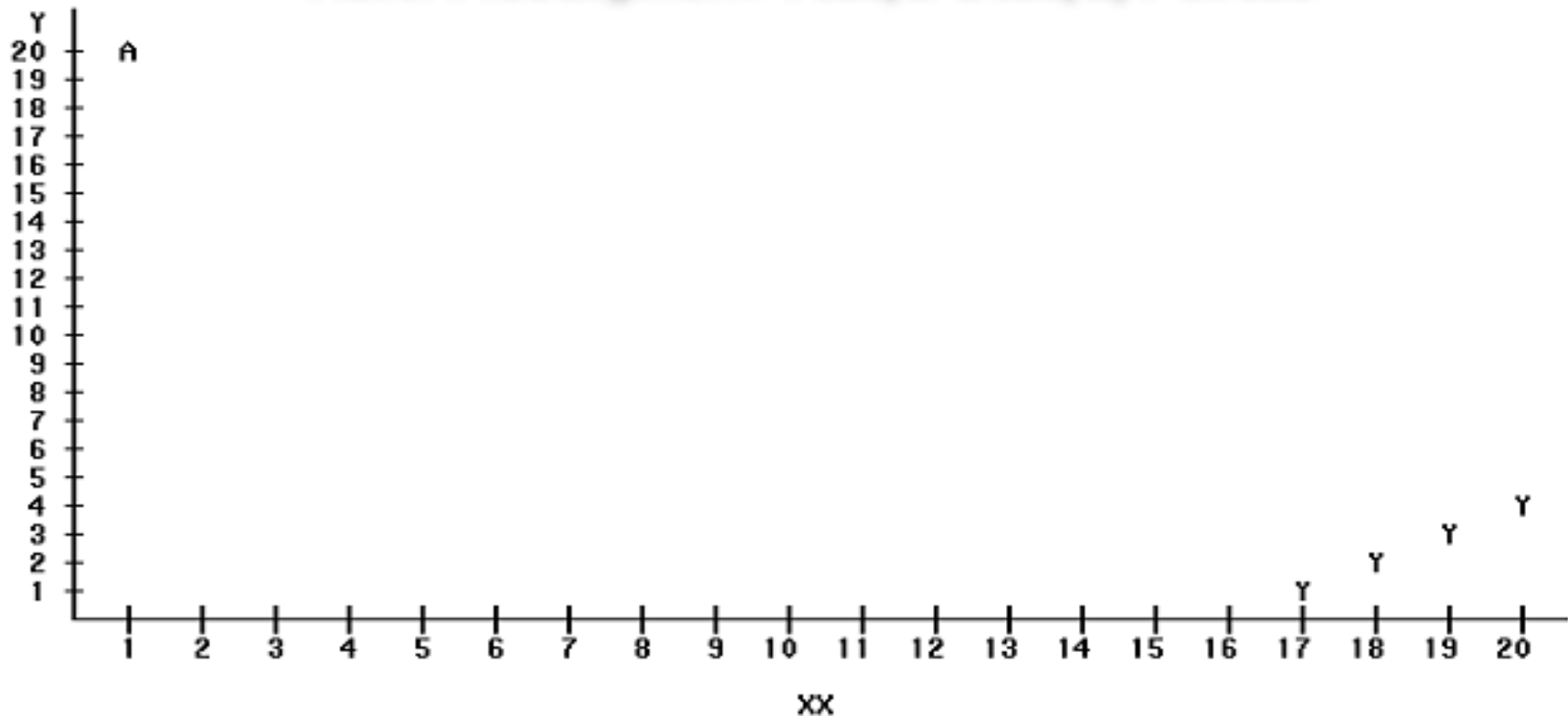


Figure 1. Plot of Y and XX.

Perfect Correlation Coefficient ... If Not for a Single Outlier

- I apply the *GenIQ Model* to the dataset of 101 points. The correlation coefficient of (GenIQvar, Y) is 0.84156, in Table 1, where GenIQvar is the transformation variable of XX.
- The corresponding scatterplot, in Figure 2, shows the assumption of *linearity does holds*.

Perfect Correlation Coefficient ... If Not for a Single Outlier

- The justification of the Linearity Assumption lies in the interpretation of the four transformed mass points:
 - ▶ The four points are in a vertical trend,
 - ▶ which is viewed as variation about a "true" transformed mass point.
- Accordingly, the relationship between Y and GenIQvar can be declared as linear.
- Thus, the ***correlation coefficient*** of 0.84156 is a ***reliable*** measure of the linear relationship between GenIQvar and Y.

Perfect Correlation Coefficient ... If Not for a Single Outlier

Plot of Y*GenIQvar. Legend: A=1 obs, B=2 obs, ..., Y=25 obs

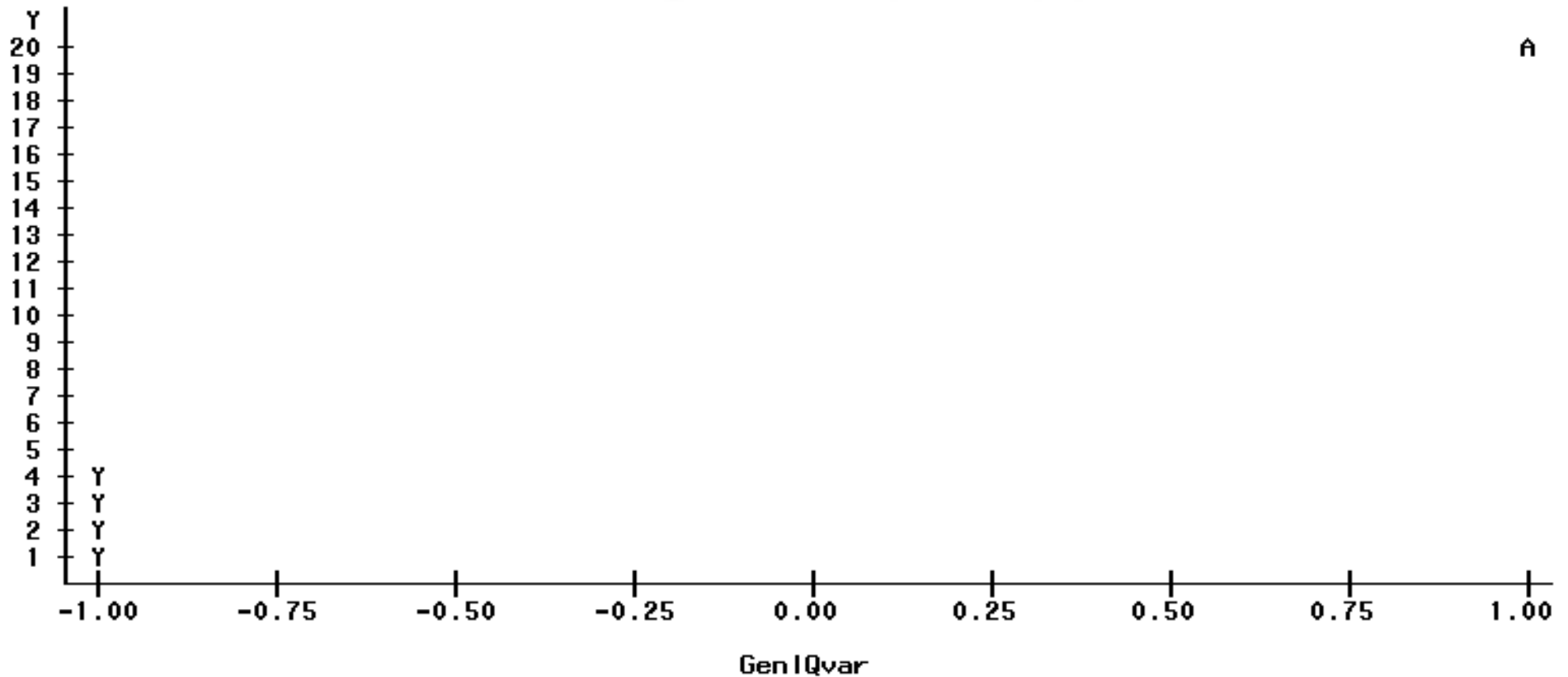


Figure 2. Plot of Y and GenIQvar.

GenIQ Model's Output

- About GenIQs output:
 - Consists of a *visual display* (called a parse tree) of the transformation of the independent variable(s), in Figure 3, and
 - the *computer code* of the transformation, in Table 2, that moderates the outlier(s) by straightening the original relationship in the data.
- In this illustration there is only one independent variable, but GenIQ can accommodate many independent variables.

GenIQ Model's Tree

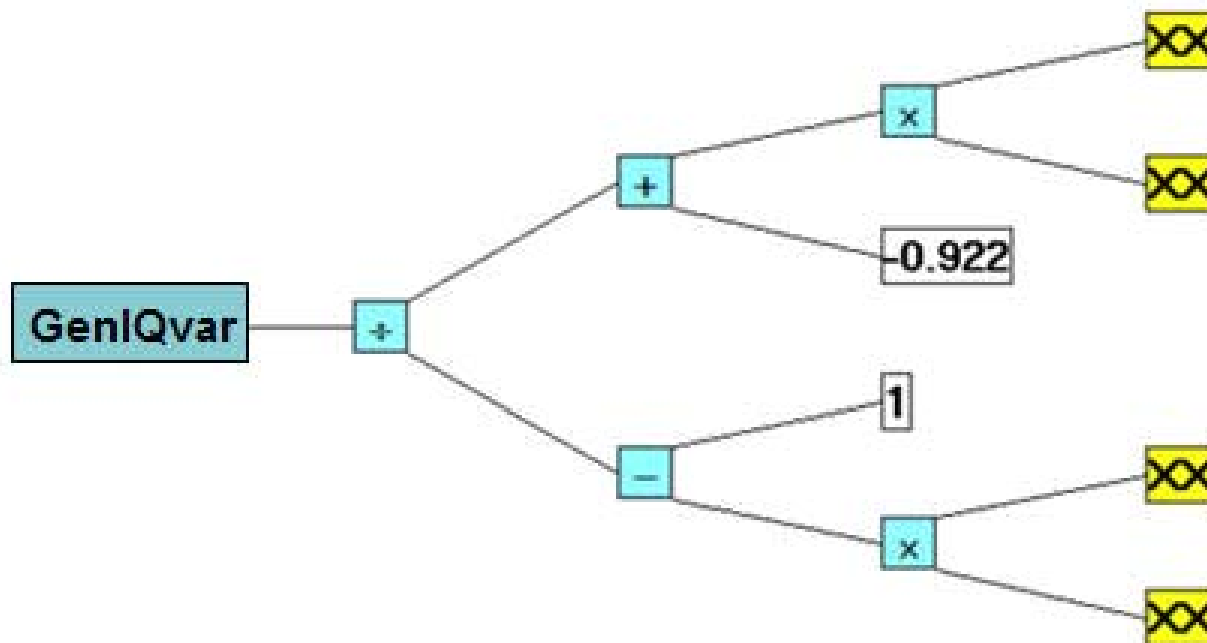


Figure 3. GenIQ Model Transformation of XX, GenIQvar.

GenIQ Model's Computer Code

Table 2. GenIQ Code for XX

```
x1 = XX;  
    x2 = XX;  
    x1 = x1 * x2;  
    x2 = 1;  
    x1 = x2 - x1;  
    x2 = -.921982;  
    x3 = XX;  
    x4 = XX;  
    x3 = x3 * x4;  
    x2 = x2 + x3;  
    If x1 NE 0 Then x1 = x2 / x1; Else x1 = 1;  
GenIQvar = x1;
```

Review of Objective

- Presented the **GenIQ Model** for moderating outliers, instead of discarding them.
- Illustrated **GenIQ** as a method for handling outliers with a simple dataset.
 - ***GenIQ works as well with all datasets: small or big, simple or complex.***
- **GenIQ** transforms the independent variable(s) to straighten a given relationship, effectively moderating the outlier(s).
- **GenIQ** is especially useful for building ordinary least squares and logistic regression models, which are sensitive to outliers.

GenIQ Model's Output

- For more about the **GenIQ Model**, go to:
<http://www.GenIQModel.com>
- This dataset comes from Huck, S. W., (2008) "Perfect Correlation Coefficient ... If Not for a Single Outlier," *STATS*, Issue 49, 9.