Power Transformations Statlet

Summary

This Statlet may be used to explore the effect of applying various power transformations to a column of numeric data. It may be used to find the transformation that makes the transformed data most closely characterized by a normal distribution. The controls on the toolbar allow the user to interactively change the power. Alternatively, the Box-Cox approach may be used to find an optimal power.

Sample StatFolio: *powerstatlet.sgp*

Sample Data

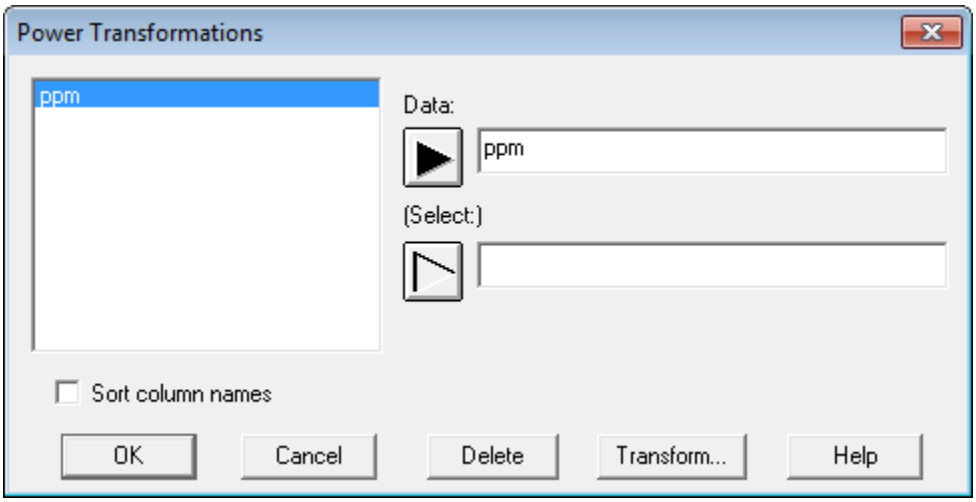
The file *groundwater.sgd* contains $n = 47$ measurements of the concentration of uranium in groundwater samples taken from a location in northwest Texas. The table below shows a partial list of the data from that file:

ppm
8.25
2.82
4.16
18.66
12.72
8.75
2.29
7.22
9.76
7.72
27.38
5.14

The concentration is measured in parts per million.

Data Input

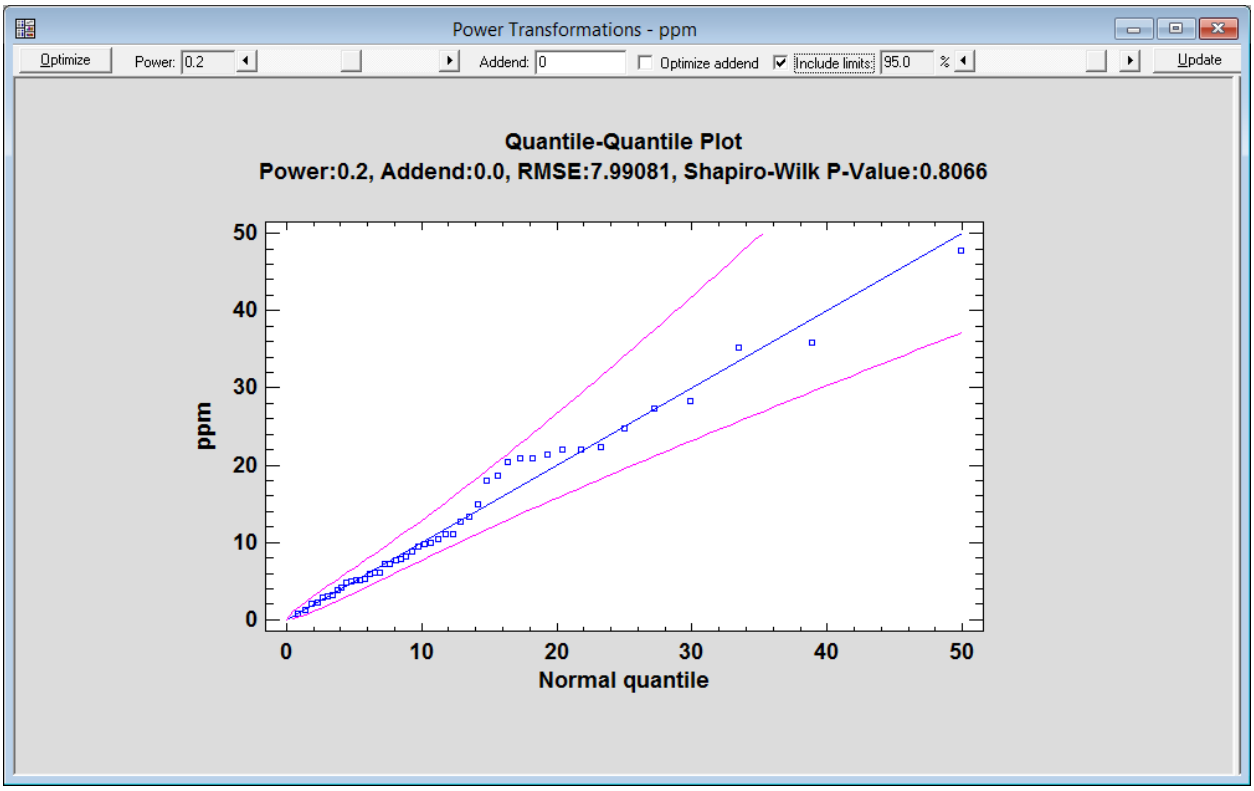
The data input dialog box requests the name of the column containing the data values to be analyzed:



- **Data:** name of the numeric column containing the *n* observations.
- **Select:** optional subset selection.

Statlet

The output of this procedure is displayed in a dynamic Statlet window:



The window shows a quantitle-quantile plot which compares the transformed values of the response variable Y to quantiles of the best-fitting normal distribution. The transformed values of Y are given by the power transformation defined by:

$$Y' = (Y + \Delta)^p \qquad \text{if } p \neq 0 \qquad (1)$$

$$Y' = \ln(Y + \Delta) \qquad \text{if } p = 0 \qquad (2)$$

where Δ is an optional addend that is added to each data value to be sure that all values are greater than 0, and p is a power between -5 and +5.

The controls on the toolbar allows you to specify the *power* and *addend*. If you press the *Optimize* button, the program will use the Box-Cox procedure to find an optimal value of p. If you first check *Optimize addend* and then press the *Optimize* button, optimal values will be found for both p and Δ.

You may also elect to *Include limits* on the plot. If so, confidence limits are added for the normal quantiles at the specified level of confidence.

Calculations

The Box-Cox procedure automatically determines the best power transformation by finding the value of *p* that minimizes the standard deviation of the observations when transformed according to the Box-Cox transformation:

$$Y' = 1 + \frac{(Y+\Delta)^p - 1}{p g^{p-1}} \qquad \text{if} \qquad p \neq 0 \tag{3}$$

$$Y' = 1 + g \ln(X+\Delta) \qquad \text{if} \qquad p = 0 \tag{4}$$

where *g* is the geometric mean of the observations after adding Δ:

$$g = \left(\prod_{i=1}^{n} (X_i + \Delta) \right)^{1/n} \tag{5}$$

The RMSE displayed by the Statlet is the sample standard deviation of the transformed values.

