Median Polish

The **Median Polish** procedure constructs a model for data contained in a two-way table. The model represents the contents of each cell in terms of:

1. a common value
2. a row effect
3. a column effect
4. a residual

Although the model used is similar to that estimated using a two-way analysis of variance, the terms in the model are estimated using medians rather than means. This makes the estimates more resistant to the possible presence of outliers.

The methodology is due to Tukey (1977). An excellent discussion is given in Velleman and Hoaglin (1981).

Sample StatFolio: *medianpolish.sgp*

Sample data:

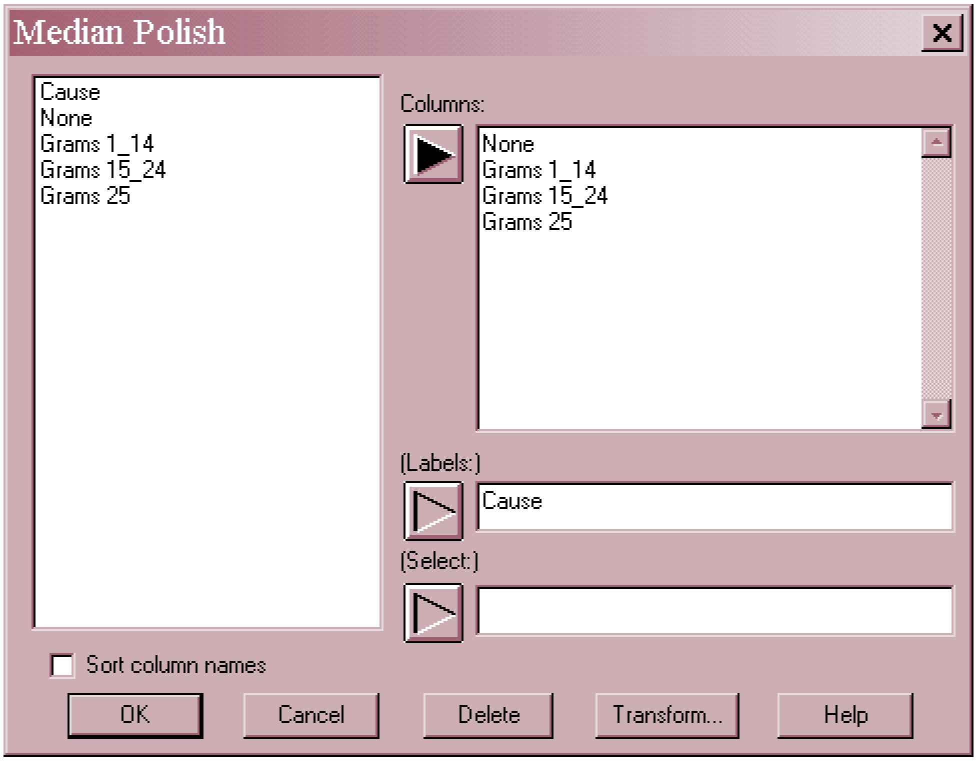
The file *tobacco.sgd* contains information on the death rates of men from various causes as a function of how much tobacco they smoked on a daily basis. The data, from Velleman and Hoaglin (1981), are shown below:

Cause	*None*	*Grams 1_14*	*Grams 15_24*	*Grams 25*
Lung cancer	0.07	0.47	0.86	1.66
Upper resp. cancer	0.00	0.13	0.09	0.21
Stomach cancer	0.41	0.36	0.10	0.31
Colon cancer	0.44	0.54	0.37	0.74
Prostrate caner	0.55	0.26	0.22	0.34
Other cancer	0.64	0.72	0.76	1.02
TB	0.00	0.16	0.18	0.29
Bronchitis	0.12	0.29	0.39	0.72
Other respitory	0.69	0.55	0.54	0.40
Thrombosis	4.22	4.64	4.60	5.99
Cardiovascular	2.23	2.15	2.47	2.25
Hemorrhage	2.01	1.94	1.86	2.33
Ulcer	0.00	0.14	0.16	0.22
Violence	0.42	0.82	0.45	0.90
Other	1.45	1.81	1.47	1.57

The data represent the number of male deaths per 1000.

Data Input

The data input dialog box specifies the columns containing the data to be analyzed.



Columns:
None
Grams 1_14
Grams 15_24
Grams 25

(Labels:)
Cause

(Select:)

- **Columns:** two or more numeric columns containing the data in each column of the table.
- **Labels:** optional labels corresponding to each row of the table.
- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* summarizes the input data and shows the original table.

Median Polish of Twoway Table

Column variables:
    None
    Grams 1_14
    Grams 15_24
    Grams 25
Number of rows: 15
Number of columns: 4

Original Table

Cause	*None*	*Grams 1_14*	*Grams 15_24*	*Grams 25*	*Row median*
Lung cancer	0.07	0.47	0.86	1.66	0.665
Upper resp. cancer	0.0	0.13	0.09	0.21	0.11
Stomach cancer	0.41	0.36	0.1	0.31	0.335
Colon cancer	0.44	0.54	0.37	0.74	0.49
Prostrate caner	0.55	0.26	0.22	0.34	0.3
Other cancer	0.64	0.72	0.76	1.02	0.74
TB	0.0	0.16	0.18	0.29	0.17
Bronchitis	0.12	0.29	0.39	0.72	0.34
Other respitory	0.69	0.55	0.54	0.4	0.545
Thrombosis	4.22	4.64	4.6	5.99	4.62
Cardiovascular	2.23	2.15	2.47	2.25	2.24
Hemorrhage	2.01	1.94	1.86	2.33	1.975
Ulcer	0.0	0.14	0.16	0.22	0.15
Violence	0.42	0.82	0.45	0.9	0.635
Other	1.45	1.81	1.47	1.57	1.52
Column median	0.44	0.54	0.45	0.74	0.54

Along the right and bottom of the table are the medians of the cells in each row and column.

Polished Table

Performing a median polish on a two-way table results in a model which expresses the data in each cell of the table as:

$$cell_{ij} = common\ value + row\ effect_i + column\ effect_j + residual_{ij} \qquad (1)$$

The *Polished Table* pane shows the fitted model:

Polished Table
Sweeping 3 times.

Cause	*None*	*Grams 1_14*	*Grams 15_24*	*Grams 25*	*Row effect*
Lung cancer	-0.5	-0.2025	0.2	0.86	0.1175
Upper resp. cancer	0.0	0.0275	0.0	-0.02	-0.4525
Stomach cancer	0.24	0.0875	-0.16	-0.09	-0.2825
Colon cancer	0.0025	0.0	-0.1575	0.0725	-0.015
Prostrate caner	0.405	0.0125	-0.015	-0.035	-0.3075
Other cancer	-0.015	-0.0375	0.015	0.135	0.2025
TB	-0.06	-0.0025	0.03	0.0	-0.3925
Bronchitis	-0.125	-0.0575	0.055	0.245	-0.2075
Other respitory	0.24	-0.0025	0.0	-0.28	-0.0025
Thrombosis	-0.305	0.0125	-0.015	1.235	4.0725
Cardiovascular	0.0925	-0.09	0.2425	-0.1175	1.685
Hemorrhage	0.0875	-0.085	-0.1525	0.1775	1.47
Ulcer	-0.0175	0.02	0.0525	-0.0275	-0.435
Violence	-0.125	0.1725	-0.185	0.125	0.0925
Other	0.035	0.2925	-0.035	-0.075	0.9625
Column effect	-0.09375	0.00875	-0.00375	0.13625	0.54625

For example, the model for the cell in the upper left corner is:

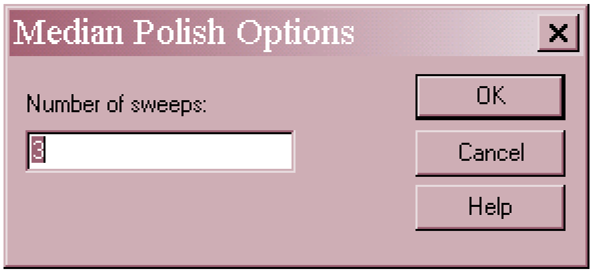
$$Cell_{1,1} = common\ value + row\ effect_1 + column\ effect_1 + residual_{1,1} \tag{2}$$

$$0.07 = 0.54625 + 0.1175 – 0.09375 – 0.5 \tag{3}$$

Intuitively, the above equation expresses the rate of lung cancer among none smokers as the sum of a constant plus a small row effect (since lung cancer has a higher rate than average), minus a small column effect (since none smokers have a lower death rate on average), plus a large residual (since none smokers rarely get lung cancer).

The column effects in the above table are of special interest. The large negative effect for none smokers indicates that they have a lower risk than the other categories. The large positive effect for the heaviest smokers indicate that they have an elevated risk. The two medium smoking categories have effects that are similar to each other.

Analysis Options

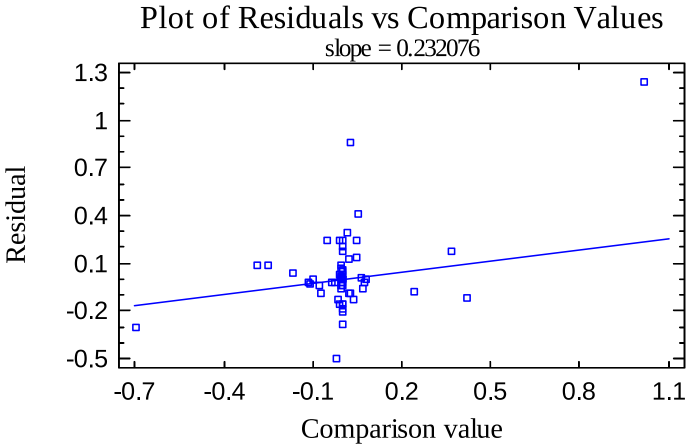

Median Polish Options

Number of sweeps: 3

- **Number of sweeps:** the number of full iterations to use in estimating the model. Since the estimation procedure is iterative, additional sweeps may make minor changes is the estimated row effects, column effects, and common value.

Residual Plot

This plot shows the residuals plotted against the comparison values:



Comparison values are calculated from:

$$c_{ij} = \frac{(roweffect_i) \times (columneffect_j)}{commonvalue} \tag{4}$$

If an additive model adequately describes the data, there should not be a discernible relationship between the residuals and the comparison values. A strong positive correlation usually indicates a need to reexpress the values in the table using some sort of power transformation.

STATGRAPHICS fits a line to the plot of residuals versus comparison values using a resistant fit (one that is not sensitive to outliers). The slope of the fitted line *b* can be using to suggest a possible power *p* to raise the table values to, where:

$$p = 1 - b \tag{5}$$

A power close to $p = 0.5$ would indicate a need to take the square roots of the data. $p = 0$ corresponds to the need to take logarithms.

In the sample data, the slope is small, suggesting that the additive model is sufficient.

Save Results

The following results can be saved to columns of the datasheet:

1. Row effects – the estimated effect of each row.
2. Column effects – the estimated effect of each column.
3. Common value – the estimated common value.
4. Residuals – the residuals for each cell of the table (the values in the polished table).
5. Comparison values – the comparison values for each cell of the table.
6. Row identifiers – the row identifier corresponding to each residual and comparison value.
7. Column identifiers – the column identifier corresponding to each residual and comparison value.