A Non-Parametric Linear Regression: Theil's Incomplete Method

Theory

Whenever the commonly used least-squares regression method is used for fitting an equation into a set of (x,y)-data points, all errors in the y-direction are normally distributed (i.e. the follow a gaussian distribution).

Non-parametric (or distribution-free) statistical methods are those, which make no assumptions about the population distribution from which the data are taken.

A simple, non-parametric approach to fit a straight line to a set of (x,y)-points is the 'Theil's incomplete method', so called to distinguish this approach from another more complex procedure (the 'complete method') developed by the same author.

Theil's 'Incomplete method' assumes that points (x_1, y_1), (x_2, y_2) . . . (x_N, y_N) are described by the equation

y = a + bx

The calculation of a and b takes place as follows:

1st step: All N data points are ranked in ascending order of x-values.

2nd step: The data are separated into two equal size (m) groups, the low (L) and the high (H) group. If N is odd the middle data point is not included to either group (hence: N = 2m or N = 2m+1).

3rd step: The slope b_i of the line connecting the i-th point of group L with i-th point of group H is calculated for all points of each group, i.e.

$$b_1 = (y_{H,1} - y_{L,1}) / (x_{H,1} - x_{L,1})$$

$$b_2 = (y_{H,2} - y_{L,2}) / (x_{H,2} - x_{L,2})$$

. . . .

$$b_m = (y_{H,m} - y_{L,m}) / (x_{H,m} - x_{L,m})$$

4th step: The median of the m slope values b_1, b_2, . . . b_m is calculated and it is taken as the best estimate of the slope (b) of the line, i.e. b = median(b_1, b_2, . . . b_m).

5th step: For each data point (x_i,y_i) the value of intercept a_i is calculated using the previously calculated slope b, i.e.

$$a_1 = y_1 - \mathbf{b}x_1$$

$$a_2 = y_2 - \mathbf{b}x_2$$

. . . .

$$a_N = y_N - \mathbf{b}x_N$$

6th step: The median of the N intercept values a_1, a_2, . . . a_N is calculated and it is taken as the best estimate of the intercept (a) of the line, i.e. a = median(a_1, a_2, . . . a_N).

The method described for the estimation of a and b has the following distinct advantages over the commonly used least-squares linear regression:

(i) It does not assume that all the errors are only in the y-direction.

(ii) It does not assume that either the x- or y-direction errors are normally distributed (i.e. it is a typical non-parametric method).

(iii) It is not affected by the presence of outlying data points (i.e. it is a 'robust method'').

Link

195.134.76.37/applets/AppletTheil/Appl_Theil2.html

A Non-Parametric Linear Regression: Theil's Incomplete Method

Theory

Whenever the commonly used least-squares regression method is used for fitting an equation into a set of (x,y)-data points, all errors in the y-direction are normally distributed (i.e. the follow a gaussian distribution).

Non-parametric (or distribution-free) statistical methods are those, which make no assumptions about the population distribution from which the data are taken.

A simple, non-parametric approach to fit a straight line to a set of (x,y)-points is the 'Theil's incomplete method', so called to distinguish this approach from another more complex procedure (the 'complete method') developed by the same author.

Theil's 'Incomplete method' assumes that points (x_1, y_1), (x_2, y_2) . . . (x_N, y_N) are described by the equation

$\mathbf{y = a + bx}$

The calculation of a and b takes place as follows:

1st step: All N data points are ranked in ascending order of x-values.

2nd step: The data are separated into two equal size (m) groups, the low (L) and the high (H) group. If N is odd the middle data point is not included to either group (hence: N = 2m or N = 2m+1).

3rd step: The slope b_i of the line connecting the i-th point of group L with i-th point of group H is calculated for all points of each group, i.e.

$$b_1 = (y_{H,1} - y_{L,1}) / (x_{H,1} - x_{L,1})$$

$$b_2 = (y_{H,2} - y_{L,2}) / (x_{H,2} - x_{L,2})$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$b_m = (y_{H,m} - y_{L,m}) / (x_{H,m} - x_{L,m})$$

4th step: The median of the m slope values b_1, b_2, . . . b_m is calculated and it is taken as the best estimate of the slope (b) of the line, i.e. b = median(b_1, b_2, . . . b_m).

5th step: For each data point (x_i,y_i) the value of intercept a_i is calculated using the previously calculated slope b, i.e.

$$a_1 = y_1 - \mathbf{b}x_1$$

$$a_2 = y_2 - \mathbf{b}x_2$$

$$\ldots\ldots\ldots$$

$$a_N = y_N - \mathbf{b}x_N$$

6th step: The median of the N intercept values a_1, a_2, . . . a_N is calculated and it is taken as the best estimate of the intercept (a) of the line, i.e. a = median(a_1, a_2, . . . a_N).

The method described for the estimation of a and b has the following distinct advantages over the commonly used least-squares linear regression:

(i) It does not assume that all the errors are only in the y-direction.

The main disadvantage of the described non-parametric method is its algorithmic nature, i.e. no specific equations are provided for the direct calculation of a and b, as in the case of least-squares regression [see Applet: Least-Squares Polynomial Approximation]. Instead, specific and repetitive steps must be made, a fact that makes manual calculations tedious. The use of a computer program (e.g. a spreadsheet) is necessary, particularly when many (x, y)-data points are involved.