DETERMINATION OF FORMAL CONFIDENCE INTERVALS OF THE REGRESSION LINES IN CASE OF LINEAR REGRESSION WITH BREAKPOINT (BP)

R.J.Oosterbaan

Used in the SegReg program (software) for segmented regression at:
https://www.waterlog.info/segreg.htm

On website https://www.waterlog.info public domain, latest upload 20-11-2017

The two regression equations are:

\[ R_{La} = Aa \left( X - AvXa \right) + AvYa \]
\[ R_{Lb} = Ab \left( X - AvXb \right) + AvYb \]

where \( Aa \) is the regression coefficient to the left of BP, \( Ab \) is the regression coefficient to the right of BP, \( X \) is a distance along the horizontal axis, \( AvXa \) is the average of the \( X \) values smaller than BP, \( AvXb \) is the average of the \( X \) values larger than BP, \( AvYa \) is the average of the \( Y \) values of the data with \( X < BP \), and \( AvYb \) is the average of the \( Y \) values of the data with \( X > BP \).

Using \( ts = \text{value of the variable in Student's distribution } ^* \) for the number of data employed at the desired confidence level, the upper confidence line to the left of BP is found from the relation:

\[ (X, Y_{1a} + ts \cdot StDevYc) \]

where:

\[ Y_{1a} = Aa \left( X - AvXa \right) + AvYa \]
\[ StDevYc = \sqrt{ s_r^2 + (X - AvXt)^2 \cdot StDevA^2 } \]

with:

\[ s_r^2 = \frac{StDevYra^2 (Na-1) + StDevYrb^2 (Nb-1)}{Nt (Nt-4)} \]

and:

\[ StDevA^2 = \frac{StDevYra^2 (Na-1) + StDevYrb^2 (Nb-1)}{(Nt-4)RedSumX^2} \]

with:

\[ RedSumX^2 = (Nt-1)(StDevX)^2 \]

where: \( AvXt \) is the average of all \( X \)-data, \( StDevYra \) is the standard deviation of the residuals of \( Y \) values after regression (or of the distances between the \( Y \) values and \( R_{La} \), \( StDevYr \)) to the left of BP, \( StDevYrb \) is the standard deviation of the residuals of \( Y \) values after regression (or of the distances between the \( Y \) values and \( R_{Lb} \), \( StDevYr \)) to the right of BP, \( Na \) is the number of data sets with \( X<BP \), \( Nb \) is the number of data sets with \( X>BP \), \( Nt \) is the total number of data sets (\( Nt = Na + Nb \)), and \( StDevX \) is the standard deviation of all the \( X \)-data (i.e. in all data sets)

Similarly, the lower confidence line to the left of BP is found from the relation:

\[ (X, Y_{1a} - ts \cdot StDevYc) \]
The upper confidence line to the right of BP is found from the relation:
\[(X, Y_1b + ts \cdot \text{StDevYc})\]
where:
\[Y_1b = Ab (X - AvXb) + AvYb\]
Similarly, the lower confidence line to the right of BP is found from the relation:
\[(X, Y_1b - ts \cdot \text{StDevYc})\]

*) [https://www.waterlog.info/t-tester.htm](https://www.waterlog.info/t-tester.htm)

**EXAMPLE**

From the example output file Dat.out (see below) we find:
Function type: 3
BP = 3.06, Aa=As=0 (for data with X<BP), Ab=Ag= - 11 (for data with X>BP),
AvXa = 1.5 (for data with X<BP), AvXb = 6.75 (for data with X>BP),
AvXt = 4.13 (for all data),
AvYa = 140 (for data with X<BP), AvYb = 99.4 (for data with X>BP),
StDevYra = 19.8 16.4 (for data with X<BP), StDevYrb = 12.0 11.4 (for data with X>BP),
Na = 12 (for data with X<BP), Nb = 12 (for data with X>BP), Nt = 24 (for all data),
StDevX = 3.04 (for all data)

Note: for “all data” see the regression without BP

With these data of Dat.out it can be calculated that:

- Eq (6): \(\text{RedSumX}^2 = (24-1) \cdot 3.04^2 = 212.6\)
- Eq (5): \(\text{StDevA}^2 = \frac{(19.8^2 \cdot (12-1) + 12.0^2 \cdot (12-1))/24}{2} = \frac{4312 + 1584}{4252} = 1.39\)
- Eq (4): \(s_r^2 = \frac{(9.14^2 + (X-4.13)^2 \cdot 1.39)}{24} = \frac{4312 + 1584}{480} = 12.3\)
- Eq (3): \(\text{StDevYc}^2 = \frac{12.3 + (X-4.13)^2 \cdot 1.39}{2} = \frac{4312 + 1584}{480} = 12.3\)

Taking for example \(X=2\) then \(\text{StDevYc} = \sqrt{\frac{9.14 + (2 - 4.13)^2 \cdot 1.39}{2}} = \sqrt{18.61} = 4.31\)

- Eq (2): \(Y_1a = 0 \cdot (X-1.5) + 140 = 140\)
Function (1): for \(X=2\) the upper confidence limit is \(140 + ts \cdot 4.31\)

Using the T-tester that can be downloaded from [www.waterlog.info/t-tester.htm](http://www.waterlog.info/t-tester.htm) we find for degrees of freedom = 24 and Probability \(P_c\) (%) = 95 that \(ts = t\)-test value \(T = 1.71\)

Hence the upper 90% confidence limit of \(Y\) where \(X=2\) is \(140 + 1.71 \cdot 4.31 = 140 + 7.37 = 147.4\)

The lower 90% confidence limit of \(Y\) where \(X=2\) is \(140 - 1.71 \cdot 4.31 = 140 - 7.37 = 132.6\)

**Note 1:** Taking \(X = BP = 3.06\) one finds from Eq (3) the value \(\text{St.Dev.Ybp} = 3.72\), being the standard deviation of \(Y\) at BP, as can be seen in the example output file Dat.out
Note 2: The standard error of the breakpoint (St.Err.BP) is found from St.Dev.Ybp as follows:
- Type 3 segmented regression: St.Err.BP = abs(St.Dev.Ybp / Ab)
  In this example St.Err.BP = 3.72 / 11 = 0.338
- Type 4 segmented regression: St.Err.BP = abs(St.Dev.Ybp / Aa)
- Type 2 segm. regression:
  St.Err.BP = 0.5*(abs(St.Dev.Ybp / Ab) + abs(St.Dev.Ybp / Aa))
Using Student’s t value one can make a confidence interval for BP using St.Err.BP

Note 3: In the example, the upper and lower confidence limits span a 90% confidence interval, with 5% probability of exceedance (or 95% cumulative probability Pc) and 5% probability of non-exceedance (see figures below).
Calculated interval probability $P(L < X < U)$ for given $L$ and $U$ values

Degrees of freedom = 24

Lower limit of $X$: $L = -1.71$
Upper limit of $X$: $U = 1.71$
Interval probability = green area = 0.900

$t$-Tester program at www.waterlog.info/t-test.htm
Results of program SegReg for segmented linear regression of Y upon X. Y is the dependent variable.
There can be different types of functions ranging from type 0 to type 6 and (for types 2, 3 and 4) two methods of calculation. The types and methods are determined with the procedure of best fit. For explanations, use the symbols function in the output scroll menu.

Name of this output file:  C:\SegReg\Dat.out
Name of input file used :  C:\SegReg\Dat.inp

No first title given
No second title given
Minimum confidence % : 90

Regression of Y upon X without breakpoint (BPx=Xmin).
X is the independent variable.

The table below gives the following series of values respectively:
<table>
<thead>
<tr>
<th>Breakpoint(BPz)</th>
<th>number of data</th>
<th>Av.Y</th>
<th>Av.X</th>
</tr>
</thead>
<tbody>
<tr>
<td>St.Dev.Y</td>
<td>St.Dev.Yr</td>
<td>St.Dev.X</td>
<td></td>
</tr>
</tbody>
</table>

BPx= 0.00
-6.78E+000
2.81E+001

Results of regression of Y upon X with optimal breakpoint (BPx)
The second (Z) of two independent variables is used.

The table below gives the following series of values respectively:
<table>
<thead>
<tr>
<th>Breakpoint(BPx)</th>
<th>number of data</th>
<th>Av.Y</th>
<th>Av.X</th>
</tr>
</thead>
<tbody>
<tr>
<td>St.Dev.Y</td>
<td>St.Dev.Yr</td>
<td>St.Dev.X</td>
<td></td>
</tr>
</tbody>
</table>

for the data with X-values smaller and greater than BPx followed by the function parameters.

Data with X < BPx :
BPx= 3.06
9.95E+000
1.98E+001

Data with X > BPx :
BPx= 3.06
-8.79E+000
1.91E+001

Parameters for function type 3 and method 2
<table>
<thead>
<tr>
<th>Slope &gt; BPx</th>
<th>Ybp</th>
<th>N&gt;/Nt</th>
<th>increase Yi</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.10E+001</td>
<td>1.40E+002</td>
<td>5.00E-001</td>
<td>2.03E+001</td>
</tr>
</tbody>
</table>

St.Err.Slope>BPx
St.Err.BPx
St.Err.N>/Nt
St.Err.Y1
2.17E+000
3.38E-001
1.02E-001
5.74E+000

St.Dev.Yr > BPx
St.Dev.Yr < BPx
St.Dev.Ybp
1.20E+001
1.98E+001
3.72E+000

Slope < BPx
St.Err.Slope<BPx
Exp1.Coeff.
0.00E+000
5.64E+000
6.77E-001
SUMMARY OF THE Y-X REGRESSION.

Function type : 3  - first a horizontal segment, then sloping.
Calc. method : 2
See help functions on Intro tabsheet

Optimal breakpoint of X (BPx) : 3.060E+000

There are two regression equations:
when X is smaller than BPx:  Y = AsX + Cs
when X is greater than BPx:  Y = AgX + Cg

As =  0.00E+000
Cs =  1.40E+002
Ag = -1.10E+001
Cg =  1.74E+002

<table>
<thead>
<tr>
<th>Serial</th>
<th>Yobs</th>
<th>X</th>
<th>Ycalc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.14E+002</td>
<td>0.00E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>2</td>
<td>1.34E+002</td>
<td>0.00E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>3</td>
<td>1.36E+002</td>
<td>0.00E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>4</td>
<td>1.65E+002</td>
<td>1.50E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>5</td>
<td>1.57E+002</td>
<td>1.50E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>6</td>
<td>1.26E+002</td>
<td>1.50E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>7</td>
<td>1.38E+002</td>
<td>1.50E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>8</td>
<td>1.22E+002</td>
<td>1.50E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>9</td>
<td>1.16E+002</td>
<td>1.50E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>10</td>
<td>1.77E+002</td>
<td>3.00E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>11</td>
<td>1.42E+002</td>
<td>3.00E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>12</td>
<td>1.54E+002</td>
<td>3.00E+000</td>
<td>1.40E+002</td>
</tr>
<tr>
<td>13</td>
<td>1.32E+002</td>
<td>4.50E+000</td>
<td>1.24E+002</td>
</tr>
<tr>
<td>14</td>
<td>1.12E+002</td>
<td>4.50E+000</td>
<td>1.24E+002</td>
</tr>
<tr>
<td>15</td>
<td>1.22E+002</td>
<td>4.50E+000</td>
<td>1.24E+002</td>
</tr>
<tr>
<td>16</td>
<td>8.64E+001</td>
<td>6.00E+000</td>
<td>1.08E+002</td>
</tr>
<tr>
<td>17</td>
<td>1.23E+002</td>
<td>6.00E+000</td>
<td>1.08E+002</td>
</tr>
<tr>
<td>18</td>
<td>9.43E+001</td>
<td>6.00E+000</td>
<td>1.08E+002</td>
</tr>
<tr>
<td>19</td>
<td>8.86E+001</td>
<td>7.50E+000</td>
<td>9.11E+001</td>
</tr>
<tr>
<td>20</td>
<td>8.44E+001</td>
<td>7.50E+000</td>
<td>9.11E+001</td>
</tr>
<tr>
<td>21</td>
<td>1.09E+002</td>
<td>7.50E+000</td>
<td>9.11E+001</td>
</tr>
<tr>
<td>22</td>
<td>8.55E+001</td>
<td>9.00E+000</td>
<td>7.46E+001</td>
</tr>
<tr>
<td>23</td>
<td>7.77E+001</td>
<td>9.00E+000</td>
<td>7.46E+001</td>
</tr>
<tr>
<td>24</td>
<td>7.83E+001</td>
<td>9.00E+000</td>
<td>7.46E+001</td>
</tr>
</tbody>
</table>