

Types of segmented regression and their statistical determination

R.J. Oosterbaan, November 2021

<https://www.waterlog.info>

Abstract

Regressions can be done to find the trend of the dependent (influenced) Y-value at increasing independent (causative, explanatory, predicting) X-values.

Linear regressions are done to find the trend as a straight line. For more detail, the domain of X-values may be divided into two parts and the linear regression can be done to the left and to the right of the separation point. For still more detail, the domain can be divided in more than 2 parts.

Instead of linear regressions one could also use a curved regressions.

The easiest, yet often quite indicative, method is the application of simple linear regressions using a domain split into two parts only. This is called segmented (piecewise) regression in splines.

This article describes the application principles of the last mentioned method.

Contents

1. Introduction, segmented regression types
 - 1.1 Types 0 and 1
 - 1.2 Type 2
 - 1.3 Type 3
 - 1.6 Type 5
 - 1.6 Type 6
 - 1.7 Type 7
 - 1.8 Type 8
2. Statistical tests
3. Summary and conclusions
4. On line references
5. Further reading
6. Appendix: partial regression for horizontal segments

1. Introduction, segmented regression types

When applying linear regression of the dependent (influenced) variable Y over the independent (causative, explanatory, influencing) X-variable divided in two adjacent parts, one will find two different regression lines separated by a breakpoint (BP), being the X-value that divides the domain of X-values into a left hand and right hand part.

The number of possible combinations of the characteristics of both linear regression lines is quite large, as will be demonstrated in continuation.

1.1 Types 0 and 1

A single, non-segmented, horizontal regression line may be indicated as Type 0 while a single sloping line may be nominated Type 1. A single, non-segmented, sloping line that is ascending (positive slope, sloping upwards) may be called Type 1a and a single sloping line that is descending (negative slope, sloping downwards) could be called Type 1d (*Figure 1*).

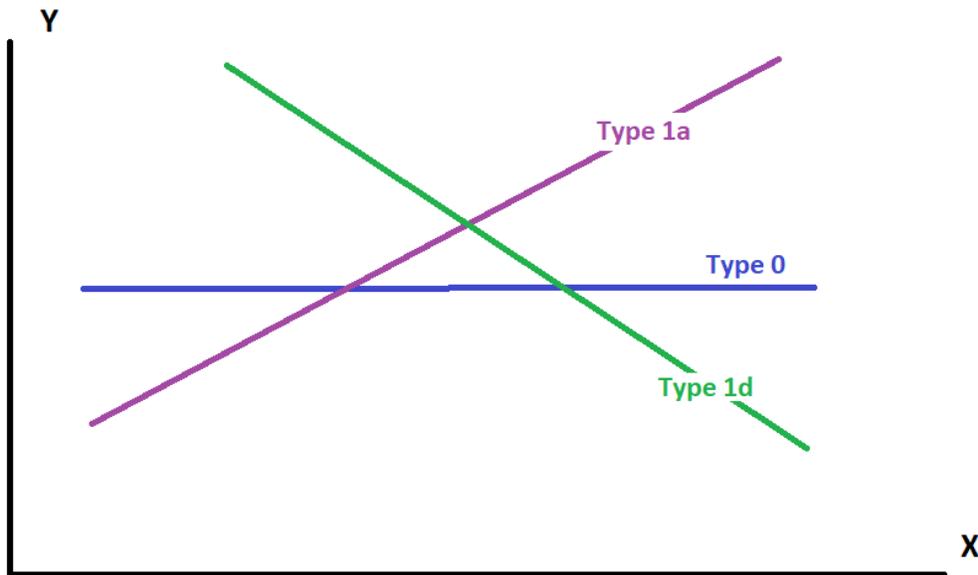


Figure 1. Examples of single regression lines. Type 0 : horizontal regression line, the regression coefficient (slope) is zero. Type 1a : upward sloping regression line (the regression coefficient is positive, greater than zero). Type 1d : downward sloping regression line (the regression coefficient is negative, less than zero).

When the regression is segmented (piecewise, in splines), there is a breakpoint (BP, being the X-value that divides the domain of X-values into a left hand and right hand part) with two different segments to the left and to the right of BP. In this case, one may discern various different segmented regression types: Type 2, 3, 4, 5, 6, 7 and 8, in dependence of the slopes of the first and second segment and the jump at the breakpoint. The slopes may be zero, positive or negative, while the jump may also be zero, positive or negative, The jump can be defined as the Y value at BP in the right hand part less the Y value at BP in the left hand part.

1.2 Type 2

Type 2 is characterized by the absence of a jump (the jump is zero) so that the regression lines to the left and to the right of BP intersect each other exactly at BP. Further, the slopes of the segments to the left and right of BP are not zero, they are either positive or negative. Examples of different Type 2 features are depicted in *Figure 2.1*.

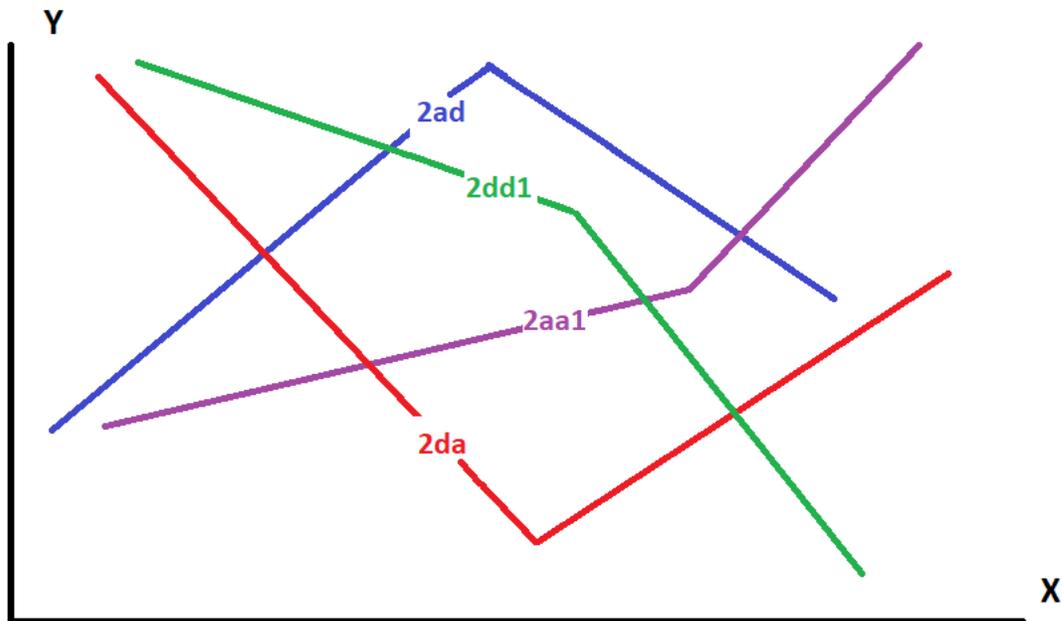


Figure 2.1. Demonstrating different features of segmented regression of Type 2. The segments to the left and to the right of the breakpoint (BP) intersect each other at BP itself. The slopes of the two different segments are either positive or negative, not zero (not horizontal).

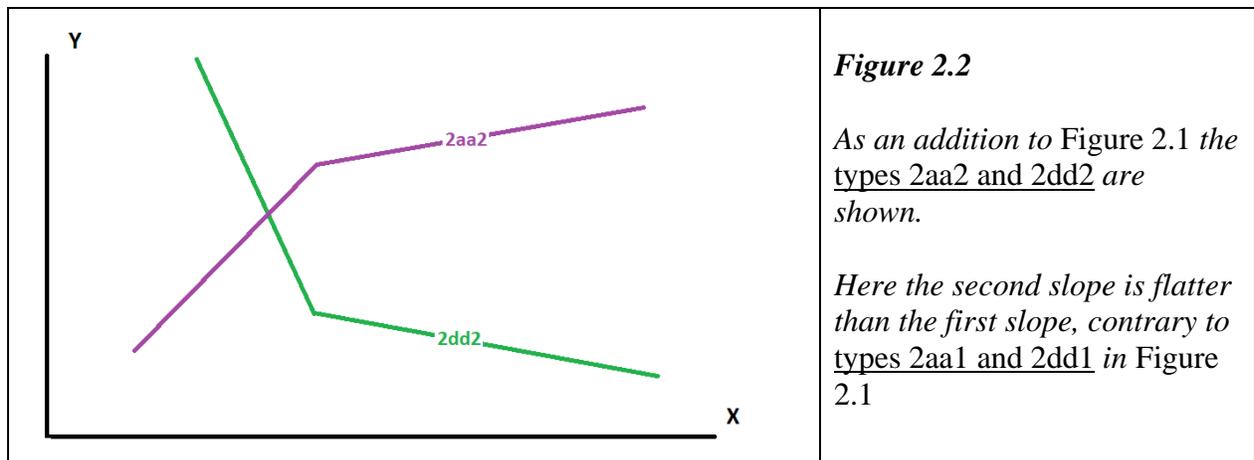
Type 2ad (blue) : the first segment slopes upward (it is ascending), while the second segment slopes downward (it is descending).

Type 2dd1 (green) : both segments slope downward (they are descending), but they are still different: the first slope is flatter than the second.

Type 2da (red) : the first segment slopes downward (it is descending), while the second segment slopes upward (it is ascending)

Type 2aa1 (purple) : both segments slope upward (they are ascending), but they are still different: the first slope is flatter than the second.

Types 2dd1 and 2aa1 can also appear reversely: the first slope is steeper than the second, yielding Type 2, subtype dd2 (Type2dd2) and Type 2, subtype aa2 (Type2aa2) respectively instead of the other way round (*Figure 2.2*).



A practical example of a Type 2 (more precisely Type 2aa1) segmented regression is presented in Figure 2.3

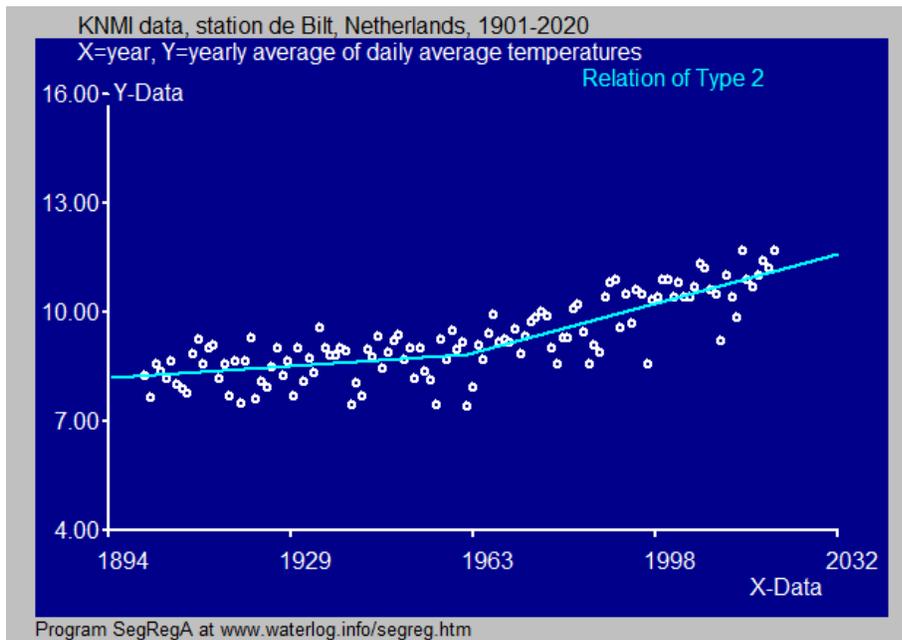
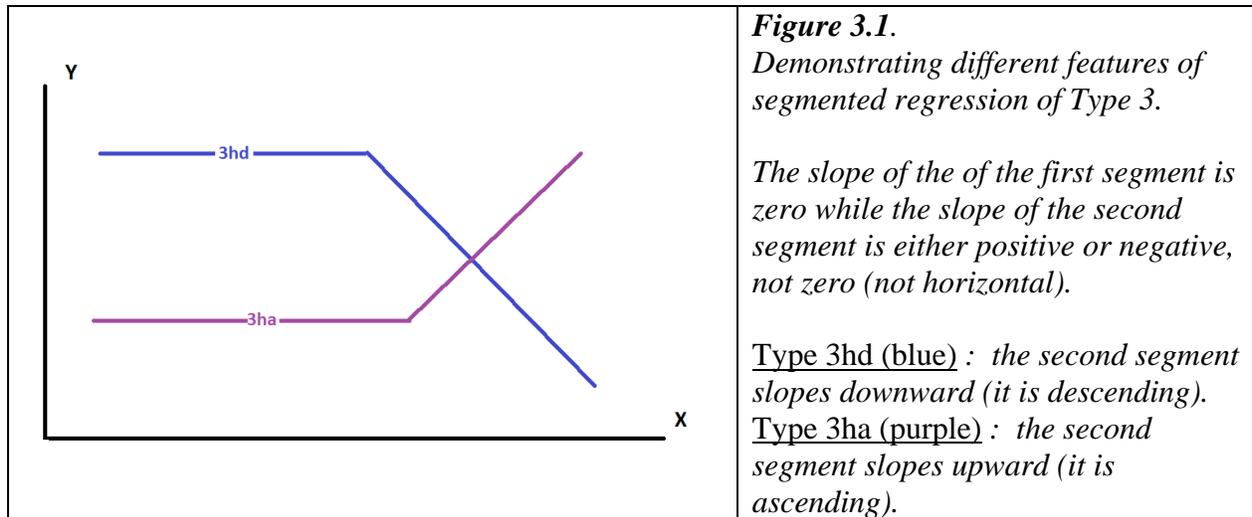


Figure 2.3. An example of a Type 2 (more precisely Type 2aa1) segmented regression. [Reference 1]. After 1963 the temperature rises more strongly.

1.3 Type 3

Type 3, like Type 2, is characterized by the absence of a jump (the jump is zero) so that the regression lines to the left and to the right of BP intersect each other exactly at BP. The difference with Type 2 is that the first slope is horizontal instead of sloping (the plateau). The correspondence is that the slope of the segments to the right of BP (the second slope) is not zero, it is either positive or negative.

Examples of different Type 3 features are depicted in *Figure 3.1*.



A practical example of a Type 3 (more precisely Type 3hd) segmented regression of the plateau type is presented in *Figure 3.2*

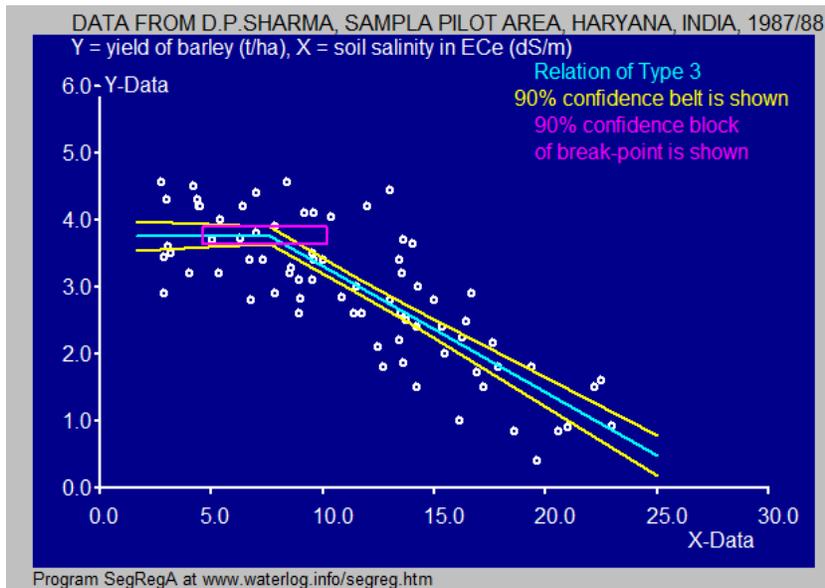
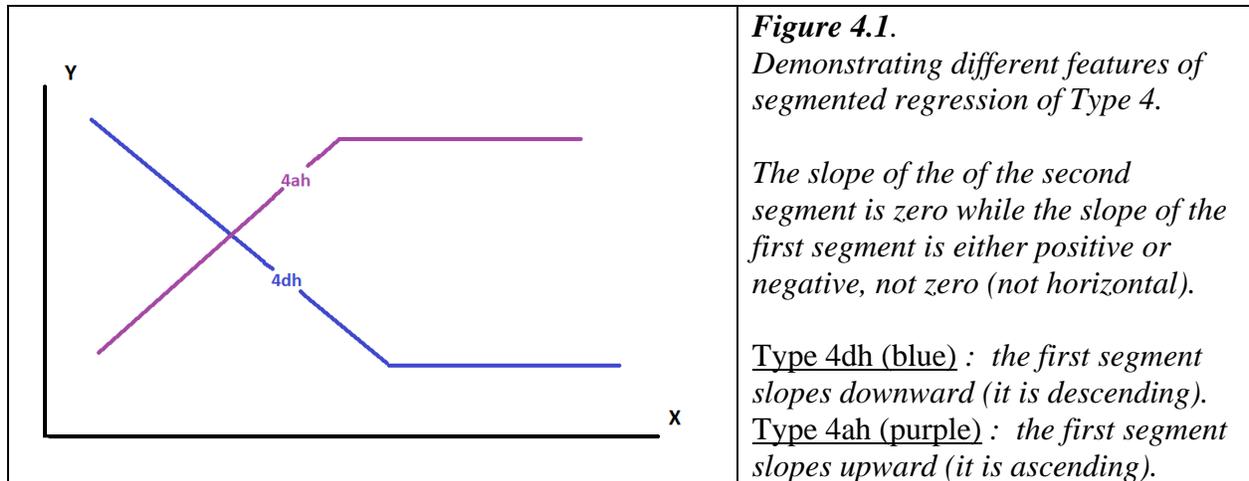


Figure 3.2. An example of a Type 3 (more precisely Type 3hd) segmented regression. [Reference 2]. When the salinity is more than 7, the yield declines. Below that there is a horizontal plateau.

1.4 Type 4

Type 4, like Types 2 and 3, is characterized by the absence of a jump (the jump is zero) so that the regression lines to the left and to the right of BP intersect each other exactly at BP. The difference with Type 2 is that the second slope is horizontal instead of sloping. The correspondence is that the slope of the segments to the left of BP (the second slope) is not zero, it is either positive or negative. The difference with Type 3 is that the second segment is horizontal instead of the first.

Examples of different Type 4 features are depicted in *Figure 4.1*.



A practical example of a Type 4 (more precisely Type 4ah) segmented regression with a plateau is presented in *Figure 4.2*

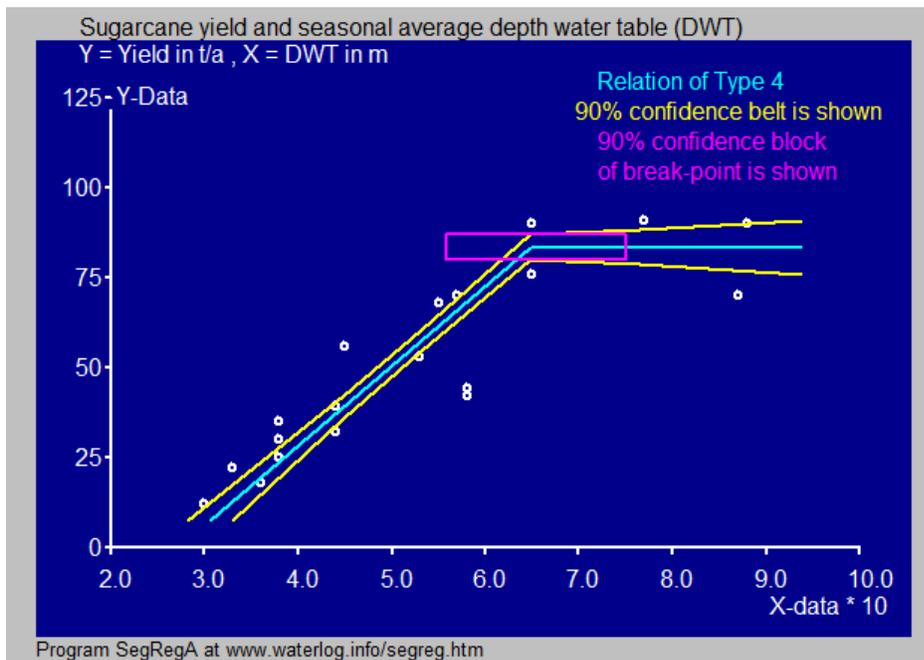
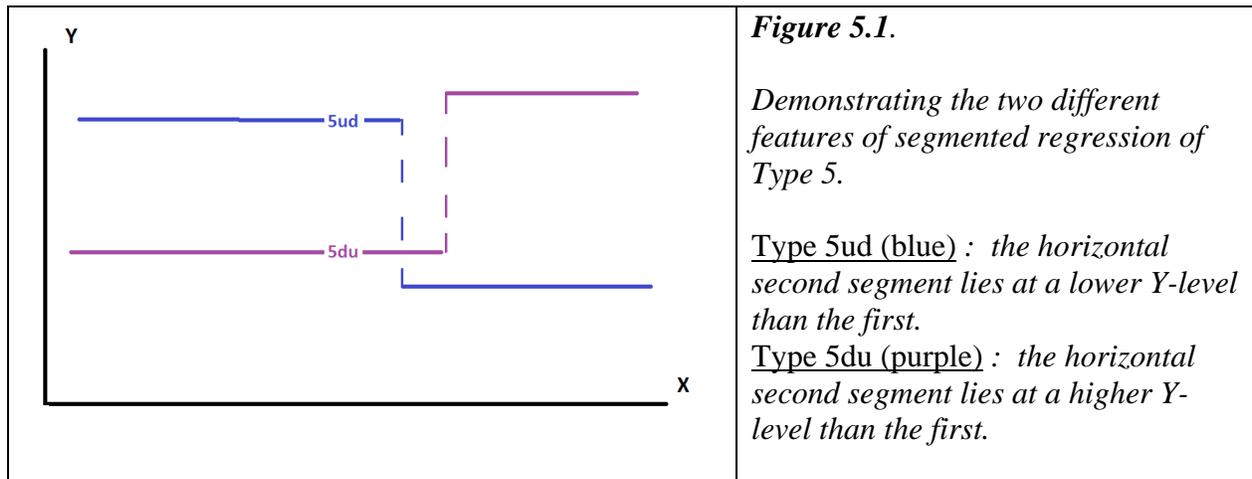


Figure 4.2. An example of a Type 4ah segmented regression. [Reference 3]. When the water table is shallower than 6.5 dm, the yield declines. After that there is a horizontal plateau.

1.5 Type 5

Type 5 has two horizontal segments. The two segments show a statistically significant jump at BP.

Examples of the two different Type 5 features are depicted in *Figure 5.1*.



A practical example of a Type 5 (more precisely Type 5ud) segmented regression is presented in *Figure 5.2*

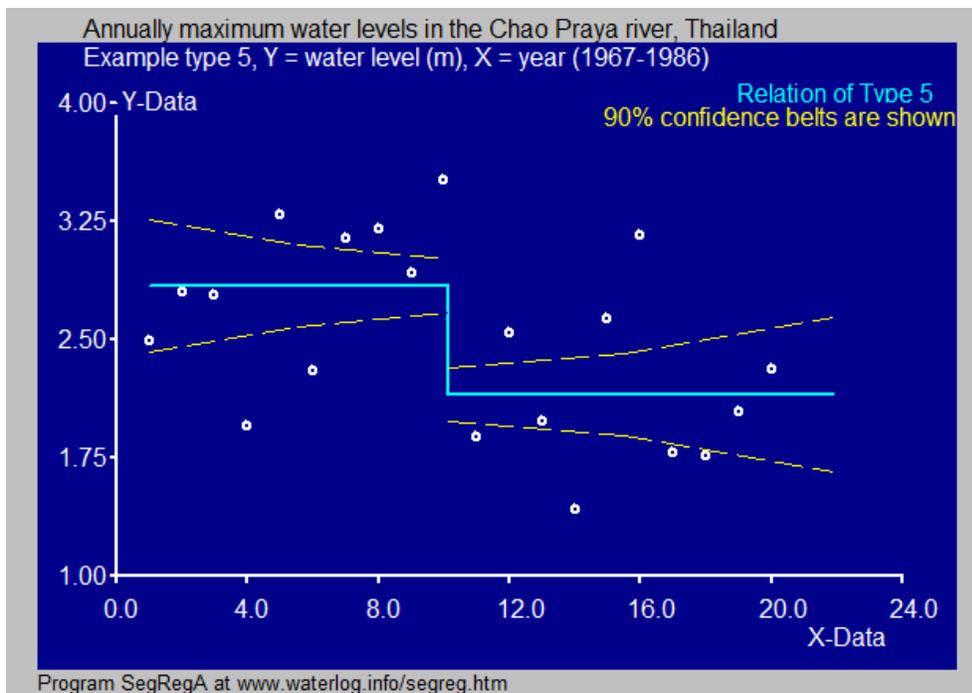
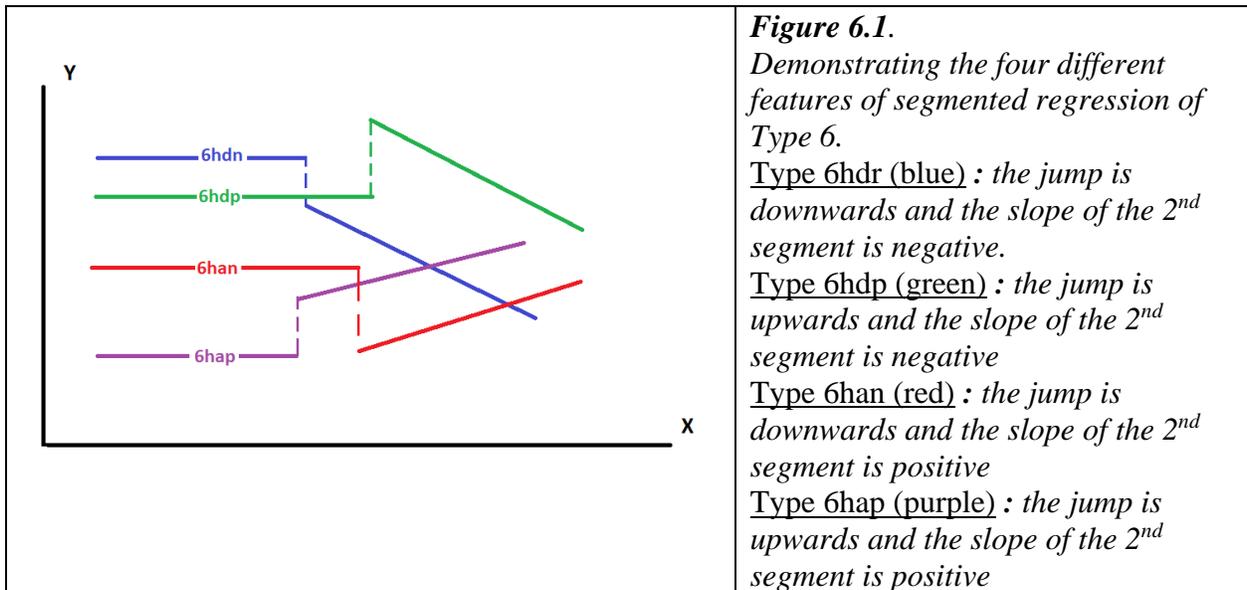


Figure 5.2. An example of a Type 5ud segmented regression. After year 10 (1976, completion of the dam)) the average water level in the river is lower than before.

1.6 Type 6

The segmented regression Type 6 and its subtypes are similar to Type 3 and its subtypes with the difference that there is a jump at the breakpoint BP. Type 3 has a first segment with zero slope (it is horizontal) and type 6 is likewise. For Type 3, the 1st and 2nd segments intersect each other at the X-value equal to BP, but for Type 6 the Y-value of the 1st segment at BP is significantly different from the Y-value of the 2nd segment at BP (*Figure 1.6*).



A practical example of a Type 6 (more precisely Type 6hap) segmented regression with a plateau is presented in *Figure 6.2*

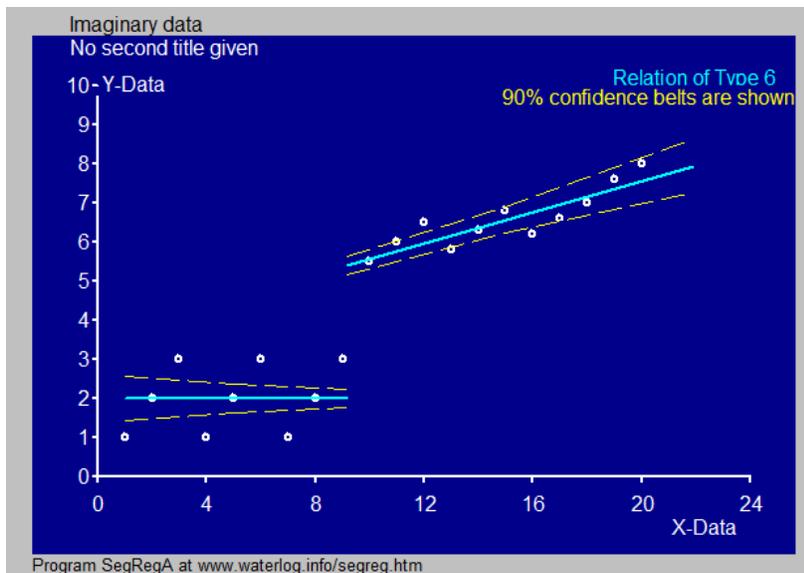
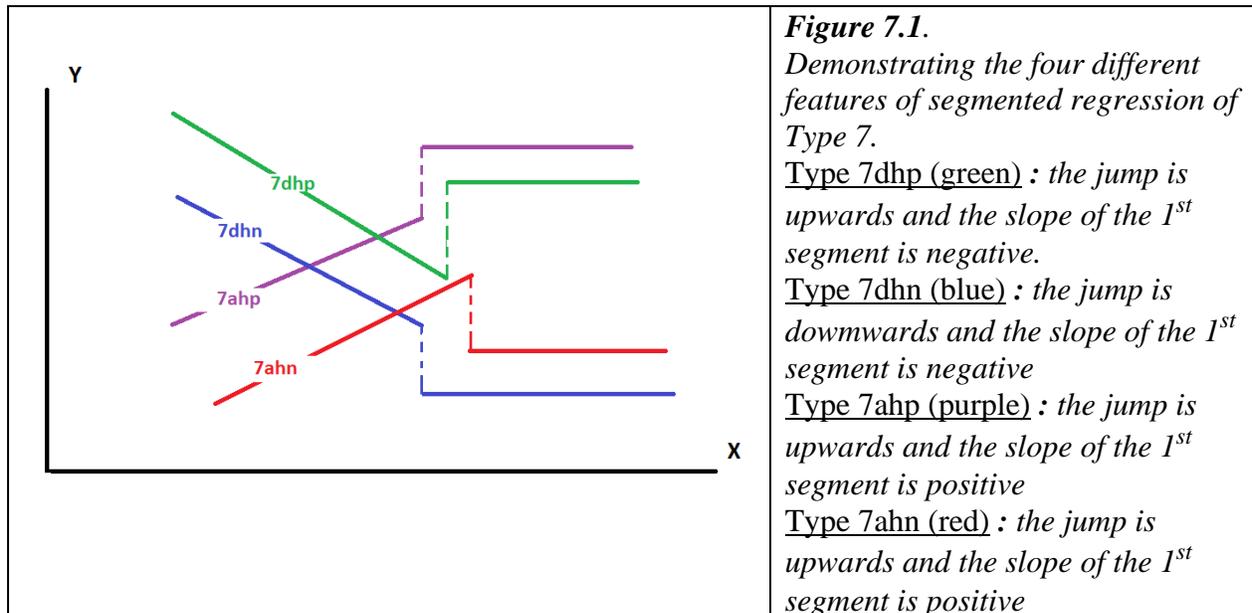


Figure 6.2. An example of a Type 6hap segmented regression. The initial trend is horizontal (the plateau). At $X=10$ there is a positive jump of Y-data followed by an upward sloping trend.

1.7 Type 7

The segmented regression Type 7 and its subtypes are similar to Type 4 and its subtypes with the difference that there is a jump at the breakpoint BP. Type 4 has a second segment with zero slope (it is horizontal) and Type 7 is likewise. For Type 4, the 1st and 2nd segments intersect each other at the X-value equal to BP, but for Type 7 the Y-value of the 1st segment at BP is significantly different from the Y-value of the 2nd segment at BP (*Figure 1.7*).



A practical example of a Type 7 (more precisely Type 7ahn) segmented regression with a plateau is presented in *Figure 7.2*. In fact, this example looks a lot like Type 4, which may be preferable.

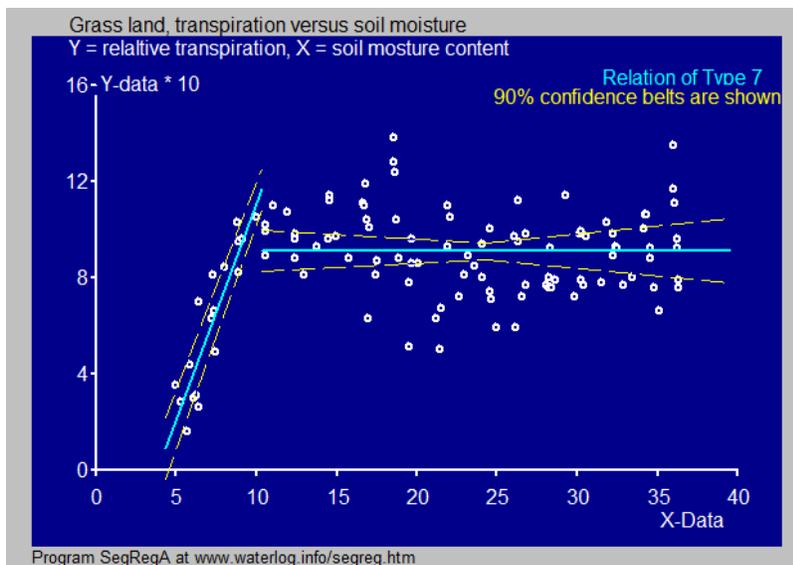


Figure 7.2. An example of a Type 7ahn segmented regression. The initial trend is sloping upward. At X=10 there is a negative jump of Y-data followed by horizontal line (the plateau).

1.8 Type 8

When none of the previous types 0 to 7 is applicable, the segmented regression type is called Type 8.

A practical example of a Type 8 segmented regression is presented in *Figure 8.1*

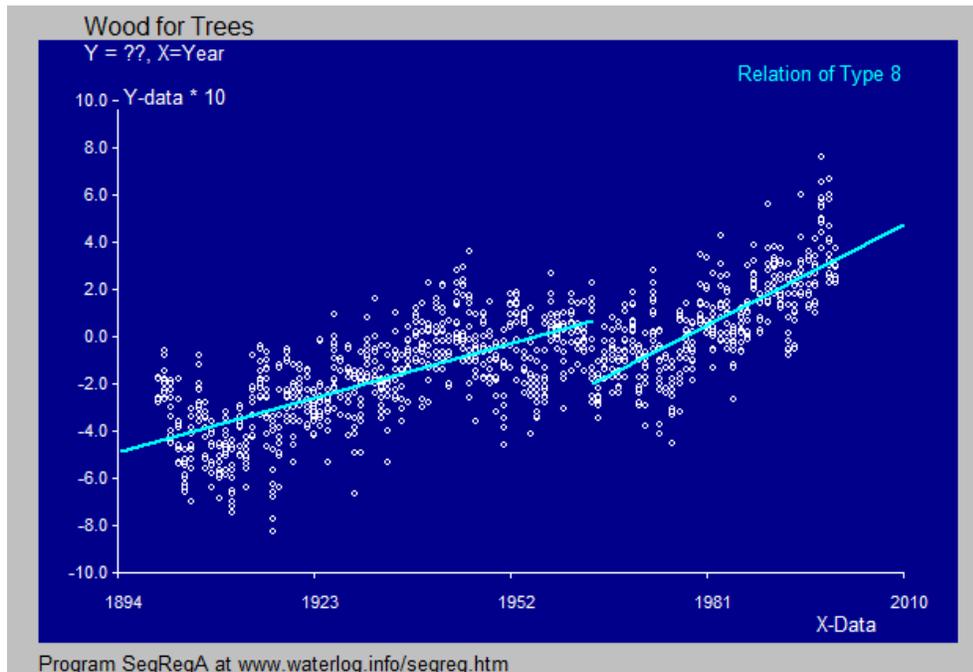


Figure 8.1. An example of a Type 8 segmented regression. There is a jump (hence Type 0, 1, 2, 3 and 4 are not applicable) and there is no horizontal trend line (meaning that Type 5, 6 and 7 are not relevant)..

2. Statistical tests

In order to perform statistical tests for the determination of the appropriate segmented regression type, one will first need to determine the following parameters:

- The total number of data pairs (N_t)
- The mean value of all X-data (X_{avt})
- The mean value of all Y-data (Y_{avt})

With these values one can find the regression coefficient or slope of the regression line (R_t) and the equation for the linear regression line for all data [*Reference 4*].

Further one needs to find:

- The standard deviation of all X-data (StdX_t)
- The standard deviation of all Y-data (StdY_t)
- The covariance of all X and Y data (Cov_t)
- The sum of squares of the differences between observed Y-values and those found by the linear regression of all data (SsqD_t).

With these values one can calculate the standard error of R_t (SteR_t) and the confidence belt of the regression line [*Reference 4*].

Next, one will first need assume a range of BP values and to determine the following statistical parameters at each BP:

- The number of data pairs with X-Value < BP (N_1)
- The number of data pairs with X-Value > BP (N_2)
- The mean value of X-data <BP (X_{av1})
- The mean value of Y-data with $X < BP$ (Y_{av1})
- The mean value of X-data >BP (X_{av2})
- The mean value of Y-data with $X > BP$ (Y_{av2})

With these values one can find the regression coefficient or slope of the regression line below $X=BP$ (R_1) and above $X=BP$ (R_2) as well as the equation for the linear regression line with $X < BP$ and $X > BP$ respectively [*Reference 4*]. Hence, one can compute from the regression equation the Y-values for each X-value less than BP and each X-value greater than BP. They will be called Y_{x1} and Y_{x2} respectively. Moreover one can find the value of Y at BP (Y_{bp1}) in the first segment where $X < BP$ and in the second segment (Y_{bp2}) where $X > BP$ [*Reference 4*].

In continuation one needs to find:

- The standard deviation of X-data below BP (StdX_1)
- The standard deviation of X-data above BP (StdX_2)
- The standard deviation of Y-data with $X < BP$ (StdY_1)
- The standard deviation of Y-data with $X > BP$ (StdY_2)
- The covariance of the X and Y data with $X < BP$ (Cov_1)
- The covariance of the X and Y data with $X > BP$ (Cov_2)
- The sum of squares of the differences between observed Y-values and those found by the linear regression of the data with $X < BP$ (SsqD_1).
- The sum of squares of the differences between observed Y-values and those found by the linear regression of the data with $X > BP$ (SsqD_2).

With this information one find the standard error of Y_{bp1} and of Y_{bp2} and the respective confidence intervals yielding the probably highest (Y_{bp1u}) and the lowest confidence value (Y_{bp1d}) of Y_{bp1} , as well as the probably highest (Y_{bp2u}) and the lowest confidence value (Y_{bp2d}) of Y_{bp2} . Here one needs Student's t-test [*Reference 5*].

Now it is possible to determine whether Y_{bp1} and Y_{bp2} are significantly from each other or not.

If not, one can join these two values and decide that there is no jump between Y_{bp1} and Y_{bp2} , so that $Y_{bp1}=Y_{bp2}$. Hence the segmented regression types 2, 3 and 4 can be used, otherwise the types 5, 6, 7 and 8 are in the game.

It still remains to be seen if one of the segments is horizontal. This can be found out employing the regression coefficients $R1$ and $R2$ together with their standard errors and Student's t -test [Reference 5] to determine the statistical significance of the coefficients. If not significant, the $R1$ and/or $R2$ can be taken equal to zero. If $R1$ is zero than Type 3 or Type 6 could be applicable, depending on the presence of a significant jump. If $R2$ is zero than Type 4 or Type 7 could be applicable, depending on the presence of a jump. When both are zero than either Type 0 or Type 5 may be relevant, also depending on the jump.

Now the most cumbersome calculations need to be undertaken. With the assumed range of breakpoints and the various regression types the differences between the observed Y -values from the, by segmented regression, computed Y -values (Y_{x1} and Y_{x2}) must be calculated. The sum of squared values of these differences should be added for each assumed BP value and for each assumed segmented regression type, giving. These sums can be called $SsqDti$ where the suffix i should indicate the case at hand. At the end of all this, one can select the lowest $SsqDti$ value of all and accept the corresponding BP and Type as the final choice.

In this fashion one has obtained the best fitting regression type in combination with the optimal BP value,

The statistical confidence interval of BP can be assessed applying the method discussed in Reference 6.

3. Summary and conclusions

In this paper, it is discussed how numerous segmented regressions are performed, each with a different BP value and for a different regression type ranging from 0 to 8. The aim is to detect the best fitting regression type in combination with the optimal BP value. To reach this goal, the method of minimization of the sum of squares of the differences between the observed Y -values and the computed ones is used.

The amount of calculation work in this procedure is so large that software in this field would be a welcome solution. The free SegReg software [Reference 7] could be an outcome.

4. On line references

Reference 1.

Applying SegRegA to the annual average temperature trend from 1900 to 2020 in the Netherlands resulting from global warming; analysis by segmented linear regression types and curved functions such as S-curve, Power curve, generalized quadratic and cubic regressions. On line: [https://www.waterlog.info/pdf/average temperature.pdf](https://www.waterlog.info/pdf/average%20temperature.pdf)
or: [Trend of annual averages of daily average temperatures in the Netherlands since 1900 first showing slow and then fast increases](#)

Reference 2.

Crop Production and Soil Salinity: Evaluation of Field Data from India by Segmented Linear Regression with Breakpoint.
Paper published in Proceedings of the Symposium on Land Drainage for Salinity Control in Arid and Semi-Arid Regions, February 25th to March 2nd, 1990, Cairo, Egypt, Vol. 3, Session V, p. 373 – 38.
On line: <https://www.waterlog.info/pdf/segmregr.pdf>
or: [CROP PRODUCTION AND SOIL SALINITY: EVALUATION OF FIELD DATA FROM INDIA BY SEGMENTED LINEAR REGRESSION WITH BREAKPOINT](#)

Reference 3.

Crop Yield and Depth of Water Table, Statistical Analysis of Data Measured in Farm Lands. On line: [https://www.waterlog.info/pdf/Crop yield and depth of water table.pdf](https://www.waterlog.info/pdf/Crop%20yield%20and%20depth%20of%20water%20table.pdf)
or: [Crop yield and depth of water table, statistical analysis of data measured in farm lands](#)

Reference 4.

Frequency and Regression Analysis of Hydrologic Data, Part II: Regression analysis. Chapter 6 in: H.P.Ritzema (Ed.), *Drainage Principles and Applications*, Publication 16, second revised edition, 1994, International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. ISBN 90 70754 3 39
On line: <https://www.waterlog.info/pdf/regtxt.pdf>
or: [6 FREQUENCY AND REGRESSION ANALYSIS OF HYDROLOGIC DATA PART II: REGRESSION ANALYSIS](#)

Reference 5.

Free software for Student's t-test at: <https://www.waterlog.info/t-tester.htm>

Reference 6.

Standard Error of the Breakpoint for Type 2 in SegReg.
On line:: <https://www.waterlog.info/pdf/bptype2.pdf>

Reference 7.

SegReg, free software for segmented regression.
On line: <https://www.waterlog.info/segreg.htm>

5. Further reading

5.1 - A paper on the statistical principles of segmented regression with break-point, including the determination of its confidence interval, can be inspected at:

<https://www.waterlog.info/pdf/segmregr.pdf>

or:

[CROP PRODUCTION AND SOIL SALINITY: EVALUATION OF FIELD DATA FROM INDIA BY SEGMENTED LINEAR REGRESSION WITH BREAKPOINT](#)

5.2 - The construction of confidence intervals of the regression segments separated by the breakpoint, and of the breakpoint itself, is described in:

<https://www.waterlog.info/pdf/confidence.pdf>

5.3 - A lecture note on statistical analysis with examples of SegReg program applications is found in this document:

<https://www.waterlog.info/pdf/analysis.pdf>

or:

[Drainage research in farmers' fields: analysis of data](#)

5.4 - Statistical significance of segmented linear regression with break-point using variance analysis and F-tests. On line: <https://www.waterlog.info/pdf/ANOVA.pdf>

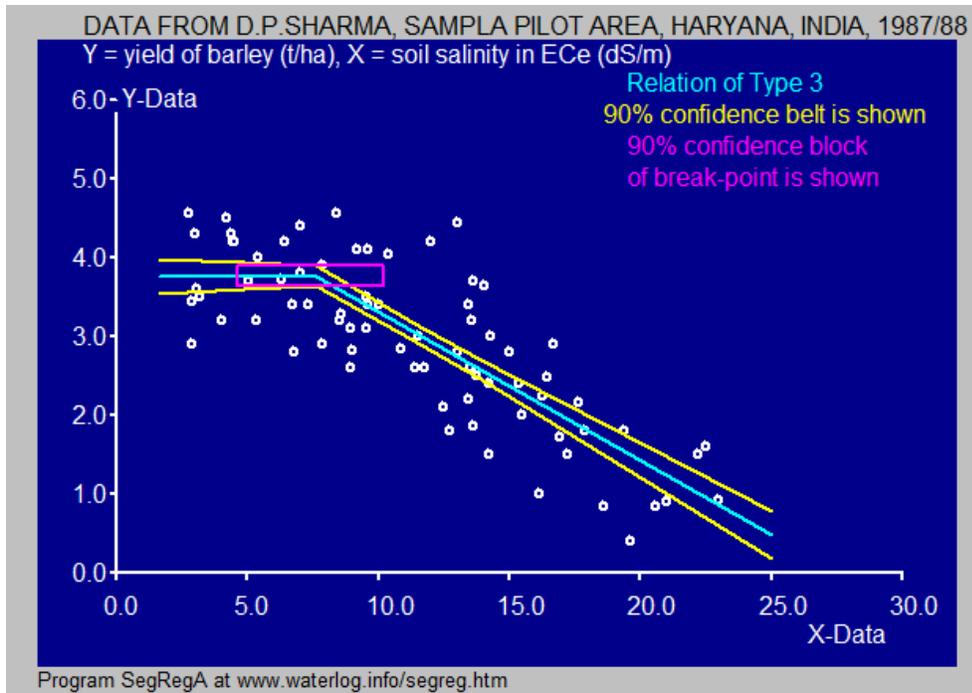
or:

[Statistical significance of segmented linear regression with break-point using variance analysis and F-tests](#)

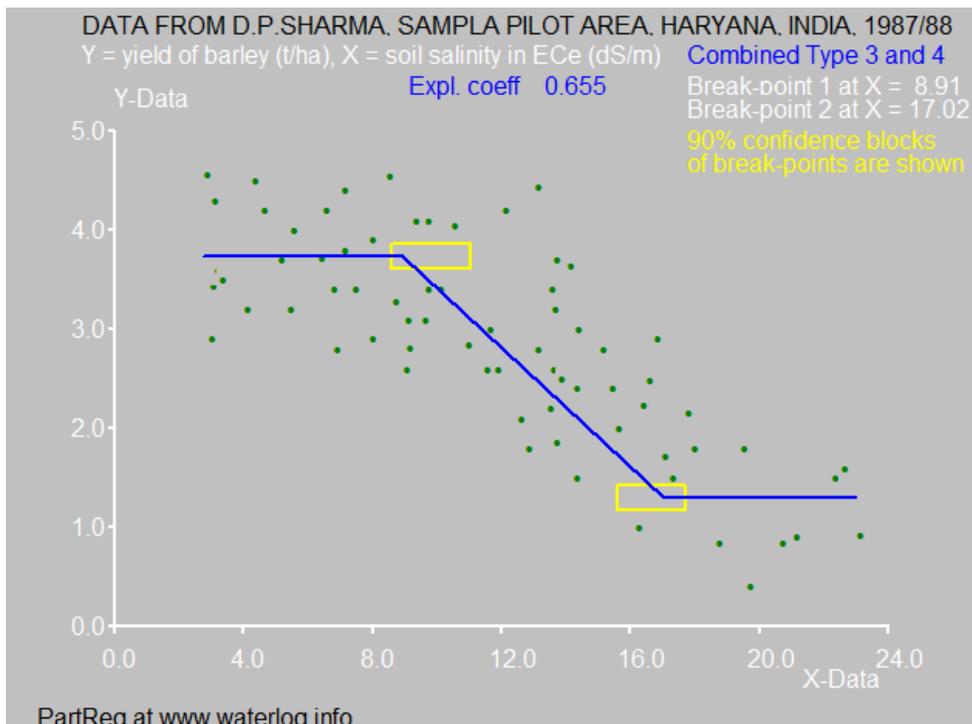
6: Appendix: Partial Regression for Horizontal Segments

Using the free PartReg software (<https://www.waterlog.info/partreg.htm>) one can find the longest horizontal stretches in an Y – X relation.

As an example, below one can see a copy of *Figure 3.2* followed by a picture showing a graph of the same data as produced by PartReg.



Copy of Figure 3.2 produced by SegReg. The breakpoint (BP) is at $X=7.6$



The same data analyzed with PartReg. The breakpoint (BP) is higher namely at $X=8.9$
 At the tail end there is another horizontal stretch.

The explanation of the difference is as follows.

The SegReg program uses the method of minimization of the sum of squares of the differences between the observed Y-values and the computed ones to find the best fitting regression type (in this case Type 3) and the optimal BP value over the entire domain of X-values. With this method, the horizontal tail end draws BP somewhat to the left as the slope of the second segment (beyond BP) is flatter than it would be without horizontal tail. The PartReg program, to the contrary, only considers those ranges of X-values over which the trend of Y – X values can be taken horizontal. Sloping segments are not taken into account.

If one is interested in the longest possible horizontal stretches in an Y – X relation, the PartReg method is the appropriate one to detect that.