Statistical significance of segmented linear regression with break-point using variance analysis and F-tests.

R.J. Oosterbaan

Abstract
The significance of different types of segmented regression analysis as performed by the SegReg model is tested by the ANOVA method (analysis of variance) as compared to simple linear regression

Subjects:

Introduction

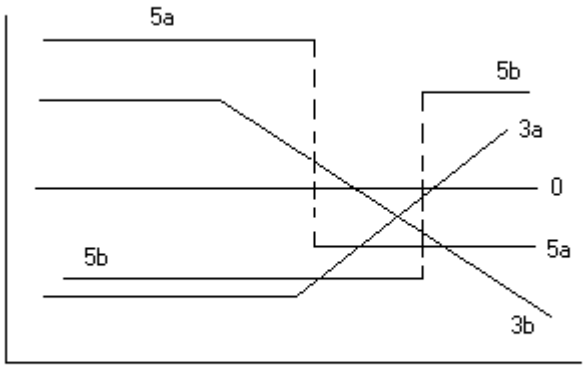
In SegReg (http://www.waterlog.info/segreg.htm) the significance of the break-point (BP) is indicated by the 90% confidence area around the BP as shown in the graphs. When the area remains within the data range, the break-point is significant, i.e. the BP gives a significant additional explanation compared to straightforward linear regression without a BP. Or, one can say that the BP analysis gives an improvement of the simple linear regression.

Although the confidence-area test makes other types of significance tests unnecessary, one may still like to perform an analysis of variance (ANOVA) and apply the F-test (named in honour of R.A. Fisher, see the reference). The following ANOVA procedure assumes that the regression is done of y (dependent variable) on x (independent variable).

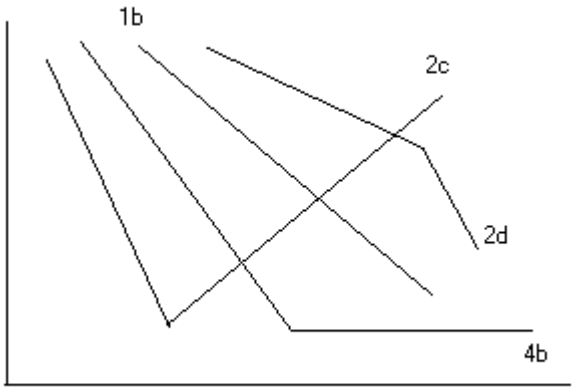
An F-test calculator is available at http://www.waterlog.info/f-test.htm

SegReg knows 6 basically different types of segmented regression as demonstrated in the figures hereunder. The analysis of variance for each type will be discussed separately.

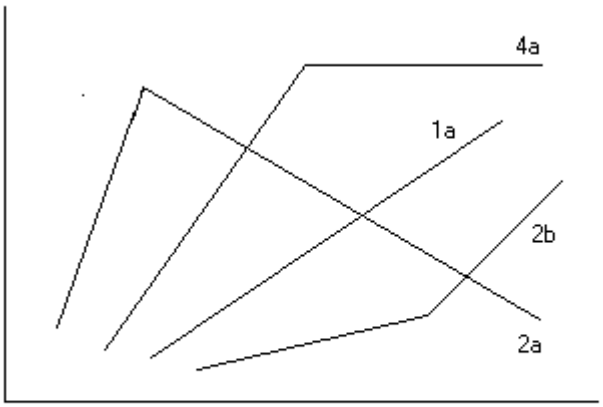
Below one can see three figures demonstrating the types of regression analysis used in SegReg. The analysis of variance for each type will be discussed separately



First part horizontal



First part sloping downward



First part sloping upward

In these figures Type 6 has not been included.

It consist of two separate sloping lines with unequal slopes and the lines are not connected.

Analysis of variance (ANOVA) for segmented linear regression with break point

The SegReg software performs segmented regressions of the dependent variable y on the independent variable x, and it produces an analysis of variance (ANOVA).

Symbols used in ANOVA

y	value of dependent variable
η	average value of y (mean)
SSD	sum of squares of deviations
r	correlation coefficient
R	overall coefficient of explanation (determination) used only when a breakpoint is present, $R = 1 - \Sigma\delta^2 / \Sigma(y-\eta)^2$, $R > r^2$ otherwise $R = r^2$
δ	residual after segmented linear regression with break-point, also called deviation from segmented linear regression
df	degrees of freedom
n	number of (x,y) data sets
x	independent variable
Var	variance or "mean square of deviation", it is the square value of the standard error (Var = SSD/df)
BP	x-value of break-point

The term $\Sigma(y-\eta)^2$ stands for "sum of squares of all reduced data", briefly "reduced sum of squares". It can be found from the SegReg output files, looking in the category of data with BP=0 (representing a linear regression of all data without break-point), using the value given for St.Dev.Y (standard deviation of y) , then multiplying it with the total number of data minus 1, and finally calculating the square value of the result:

$$\Sigma(y-\eta)^2 = [(\text{St.Dev.Y}).(n-1)]^2$$

The value of r^2 can be found directly in the same category of data. In the SegReg output it is indicated by corr.coeff.sq. (correlation coefficient squared)

Note that $r^2 = 1 - \Sigma\varepsilon^2 / \Sigma(y-\eta)^2$, where ε is the residual after linear regression, also called deviation from linear regression.

The value of $\Sigma\varepsilon^2$ is found as

$$\Sigma\varepsilon^2 = \Sigma(y- \breve{y})^2$$

where $\breve{y}$ is the expected value of y by linear regression.

The linear regression equation can be expressed as

$$\check{y} = a\,(x - \chi) + \eta$$

where a is the tangent of the slope of the regression line (also called regression coefficient), χ is the average value of x and η is the average value of y. The values of χ and η are called parameters and they influence the degrees of freedom. The value of a can be calculated from the values of x and y.

Alternatively, the regression equation can be expressed as

$$\check{y} = a\,x - a\,\chi + \eta = a\,x + b$$

In this case the value of a and $b = \eta - a\,\chi$ may be called the parameters.

In SegReg, the value of R is found in the group of results under "parameters for function type …". However, when the breakpoint is insignificant, only the regression without breakpoint is shown and the parameter R is absent, because then $R = r^2$

In the presence of a breakpoint (BP) there are two sets of data, one set (a) to the left and one set (b) to the right of it: x_a, b_b, y_a, y_b, η_a, η_b, r_a, r_b. In this case the overall average of y is represented by η_o and overall correlation coefficient r for the linear regression over the entire domain is represented by r_o

For variance analysis in SegReg the following tables are found when using the "Anova" button on the "Output" tabsheet. The tables deal with SegReg Type 1, 5, 3 and 4, 2 and 6 respectively. The crucial value of df is shown in bold.

Table 1. ANOVA table for the segmented linear regression without breakpoint **Type 1** (a sloping line)

Nr.	Description	SSD	df	Variance	F-test variable
1	total variation *) (initial, without regression)	$\Sigma(y-\eta)^2$ (SSD_1)	n-1	$SSD_1 / (n-1)$ (VAR_1)	
2	explanation by simple linear regression without BP ^)	$r^2\,\Sigma(y-\eta)^2$ (SSD_2)	1	$SSD_2 / 1 = SSD_2$ (VAR_2)	F (1, n-2) = VAR_2 / VAR_3
3	remaining unexplained after linear regression (deviations or residuals from linear regression model)	$(1-r^2)\,\Sigma(y-\eta)^2$ (SSD_3)	n-2	SSD3 / (n-2) (VAR_3)	

*) 1 df is lost for use of the mean of y

^) Another df is lost for use of the slope (regression coefficient)

Table 2. ANOVA table for the segmented linear regression with breakpoint, **Type 5**. (two horizontal lines at different levels). Two correlation coefficients are used next to r_o : r_a and r_b , and also two means next to η_o : η_a and η_b (subscript **a** stands for data to the left of the breakpoint and **b** for those to the right)

Nr.	Description	SSD	df	Variance	F-test variable
1	total variation *) (initial, without regression)	$\Sigma(y-\eta_o)^2$ ($SSD_{\mathbf{1}}$)	n-1	$SSD_{\mathbf{1}}$ / (n-1) (VAR_1)	
2	explanation by simple linear regression without BP ^)	$r_o^2\,\Sigma(y-\eta_o)^2$ ($SSD_{\mathbf{2}}$)	1	$SSD_{\mathbf{2}}$ / 1 = $SSD_{\mathbf{2}}$ (VAR_2)	F (1, n-2) = $VAR_{\mathbf{2}}$ / $VAR_{\mathbf{3}}$
3	remaining unexplained after linear regression (deviations or residuals from linear regression model)	$(1- r_o^2)\,\Sigma(y-\eta)^2$ ($SSD_{\mathbf{3}}$)	n-2	SSD3 / (n-2) (VAR_3)	
4	additional explanation by BP analysis compared to simple linear regression #)	$r_a^2\,\Sigma(y_a-\eta_a)^2$ + $r_b^2\;\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{4}}$)	1	$SSD_{\mathbf{4}}$ / 1 = $SSD_{\mathbf{4}}$ (VAR_4)	F (1, n-3) = VAR_4 / VAR_5
5	unexplained (residual) after BP analysis Type 5 (deviations from BP regression model)	$(1- r_a^2)\,\Sigma(y_a-\eta_a)^2$ + $(1- r_b^2)\,\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{5}}$)	n-3	$SSD_{\mathbf{5}}$ / (n-3) ($VAR_{\mathbf{5}}$)	
6	total explanation by segmented linear regression with BP &)	$R.\Sigma(y-\eta)^2$ ($SSD_{\mathbf{6}}$)	2	$SSD_{\mathbf{6}}$ / 2 (VAR_6)	F (2, n-3) = VAR_6 / VAR_5

*) 1 df is lost for use of the mean

^) Another df is lost for use of the slope (regression coefficient)

#) In type 5 **one** extra degree of freedom is lost owing to the use of the second mean

&) This is the explanation compared to the initial situation without regression

Table 3. ANOVA table for the segmented linear regression with breakpoint, **Types 3 and 4**. (one sloping line and one horizontal). Two correlation coefficients are used next to r_o : r_a and r_b, and also two means next to η_o : η_a and η_b (subscript **a** stands for data to the left of the breakpoint and **b** for those to the right)

Nr.	Description	SSD	df	Variance	F-test variable
1	total variation *) (initial, without regression)	$\Sigma(y-\eta)^2$ ($SSD_{\mathbf{1}}$)	n-1	$SSD_{\mathbf{1}}$ / (n-1) (VAR_1)	
2	explanation by simple linear regression without BP ^)	$r_o^2\,\Sigma(y-\eta)^2$ ($SSD_{\mathbf{2}}$)	1	$SSD_{\mathbf{2}}$ / 1 = $SSD_{\mathbf{2}}$ (VAR_2)	F (1, n-2) = $VAR_{\mathbf{2}}$ / $VAR_{\mathbf{3}}$
3	remaining unexplained after linear regression (deviations or residuals from linear regression model)	$(1- r_o^2)\,\Sigma(y-\eta)^2 + (1- r_b^2)\,\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{3}}$)	n-2	SSD3 / (n-2) (VAR_3)	
4	additional explanation by BP analysis compared to simple linear regression #)	$r_a^2\,\Sigma(y_a-\eta_a)^2 + r_b^2\,\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{4}}$)	2	$SSD_{\mathbf{4}}$ / 1 = $SSD_{\mathbf{4}}$ (VAR_4)	F (1, n-4) = VAR_4 / VAR_5
5	unexplained (residual) after BP analysis Type 3 or 4 (deviations from BP regression model)	$(1 - r_a^2)\,\Sigma(y_a-\eta_a)^2 + (1- r_b^2)\,\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{5}}$)	n-4	$SSD_{\mathbf{5}}$ / (n-4) ($VAR_{\mathbf{5}}$)	
6	total explanation by segmented linear regression with BP &)	$R.\Sigma(y-\eta)^2$ ($SSD_{\mathbf{6}}$)	2	$SSD_{\mathbf{6}}$ / 2 (VAR_6)	F (2, n-4) = VAR_6 / VAR_5

*) 1 df is lost for use of the mean

^) Another df is lost for use of the slope (regression coefficient)

#) In type 3 and 4 **two** extra degrees of freedom are lost owing to the use of the second mean and the second regression coefficient

&) This is the explanation compared to the initial situation without regression

Table 4. ANOVA table for the segmented linear regression with breakpoint, **Types 2 and 6** (one sloping line and one horizontal). Two correlation coefficients are used next to r_o : r_a and r_b, and also two means next to η_o : η_a and η_b (subscript **a** stands for data to the left of the breakpoint and **b** for those to the right)

Nr.	Description	SSD	df	Variance	F-test variable
1	total variation *) (initial, without regression)	$\Sigma(y-\eta)^2$ ($SSD_{\mathbf{1}}$)	n-1	$SSD_{\mathbf{1}}$ / (n-1) (VAR_1)	
2	explanation by simple linear regression without BP ^)	$r_o^2\,\Sigma(y-\eta)^2$ ($SSD_{\mathbf{2}}$)	1	$SSD_{\mathbf{2}}$ / 1 = $SSD_{\mathbf{2}}$ (VAR_2)	F (1, n-2) = $VAR_{\mathbf{2}}$ / $VAR_{\mathbf{3}}$
3	remaining unexplained after linear regression (deviations or residuals from linear regression model)	$(1- r_o^2)\,\Sigma(y-\eta)^2 + (1- r_b^2)\,\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{3}}$)	n-2	SSD3 / (n-3) (VAR_3)	
4	additional explanation by BP analysis compared to simple linear regression #)	$r_a^2\,\Sigma(y_a-\eta_a)^2 + r_b^2\,\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{4}}$)	3	$SSD_{\mathbf{4}}$ / 1 = $SSD_{\mathbf{4}}$ (VAR_4)	F (1, n-3) = VAR_4 / VAR_5
5	unexplained (residual) after BP analysis Type 3 or 4 (deviations from BP regression model)	$(1 - r_a^2)\,\Sigma(y_a-\eta_a)^2 + (1- r_b^2)\,\Sigma(y_b-\eta_b)^2$ ($SSD_{\mathbf{5}}$)	n-5	$SSD_{\mathbf{5}}$ / (n-5) ($VAR_{\mathbf{5}}$)	
6	total explanation by segmented linear regression with BP &)	$R.\Sigma(y-\eta)^2$ ($SSD_{\mathbf{6}}$)	3	$SSD_{\mathbf{6}}$ / 3 (VAR_6)	F (2, n-5) = VAR_6 / VAR_5

*) 1 df is lost for use of the mean

^) Another df is lost for use of the slope (regression coefficient)

#) In types 2 and 6, **three** more degrees of freedom are lost owing to the use of the second mean, and the two slopes

&) This is the explanation compared to the initial situation without regression

Examples of F-testing

If one wishes to test the significance of the simple linear regression one uses the F-statistic F(1,n-2) = Var2/Var3 (see Table 1). The F-statistic follows the F-distribution, named F in honour of R.A. Fisher. To be significant at probability level P, the F-statistic must be greater than the F-value found in F-tables at the probability P.

One may also wish to test if the additional explanation (3) could have arisen by chance. Hence the hypothesis is that BP analysis does not provide a significant extra contribution to the success of the simple linear regression. This is the "null-hypothesis". If so, Var4 and Var5 both are independent estimates of the total variance Var1. The test-statistic under the null-hypothesis for SegReg regression types 3 and 4 (see Table 3) becomes

$F_0(df_3,n-4) = F_0(2,n-4) = Var4/Var5$.

Under the null-hypothesis the F_0-value must be less than the F-value found in F-tables.

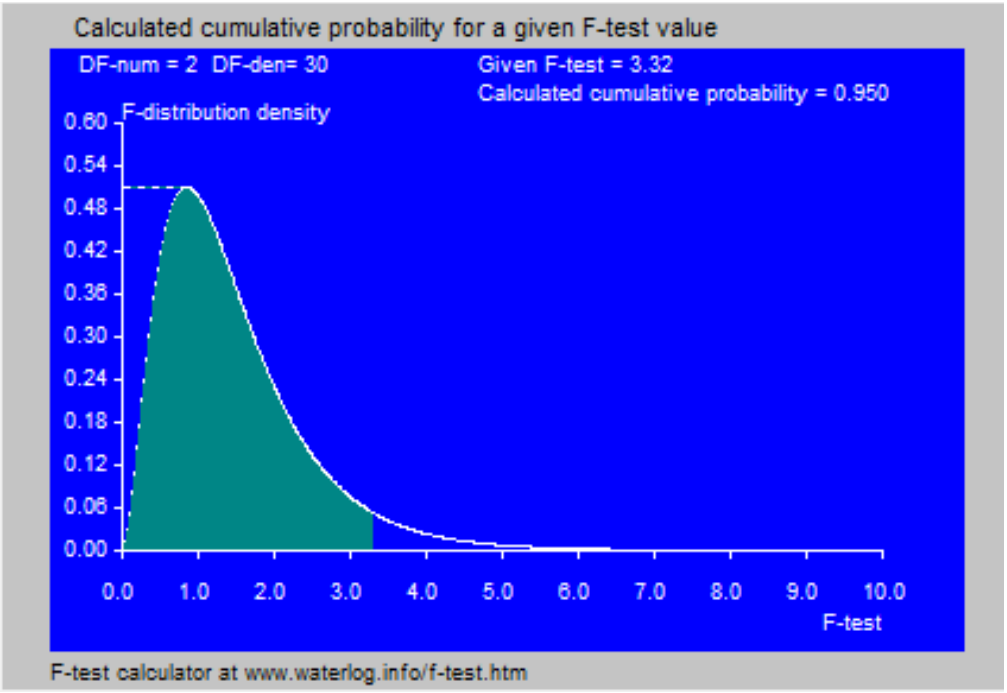
Some F-values at 5% probability of exceedance (or 95% non-exceedance, $F_{0.95}$) are:

$df_3 = 2$	$F_{0.95}$	
n-4 = 10	4.10	
n-4 = 20	3.49	
n-4 = 30	3,32	(see figure below)
n-4 = 120	3.07	
n-4 = 500	3.01	
n-4 > 1000	3.00	

Download F-test calculator from:
www.waterlog.info/exe/f-testzip.exe

Now, if $F_0 > F_{0.95}$ there is less than 5% chance that the BP analysis did not contribute significantly to the results. Taking a less than 5% risk, one may conclude that the null-hypothesis can be rejected and the conclusion is that the BP analysis has given a significant extra contribution to the regression compared to a simple linear regression (the BP is statistically "significant").

On the other hand, when Fo<F, the null-hypothesis is not rejected and the extra contribution is considered not significant, but one runs a risk of less than 5% that non-rejection of the null-hypothesis is unjustified.



Result of the F-test calculator

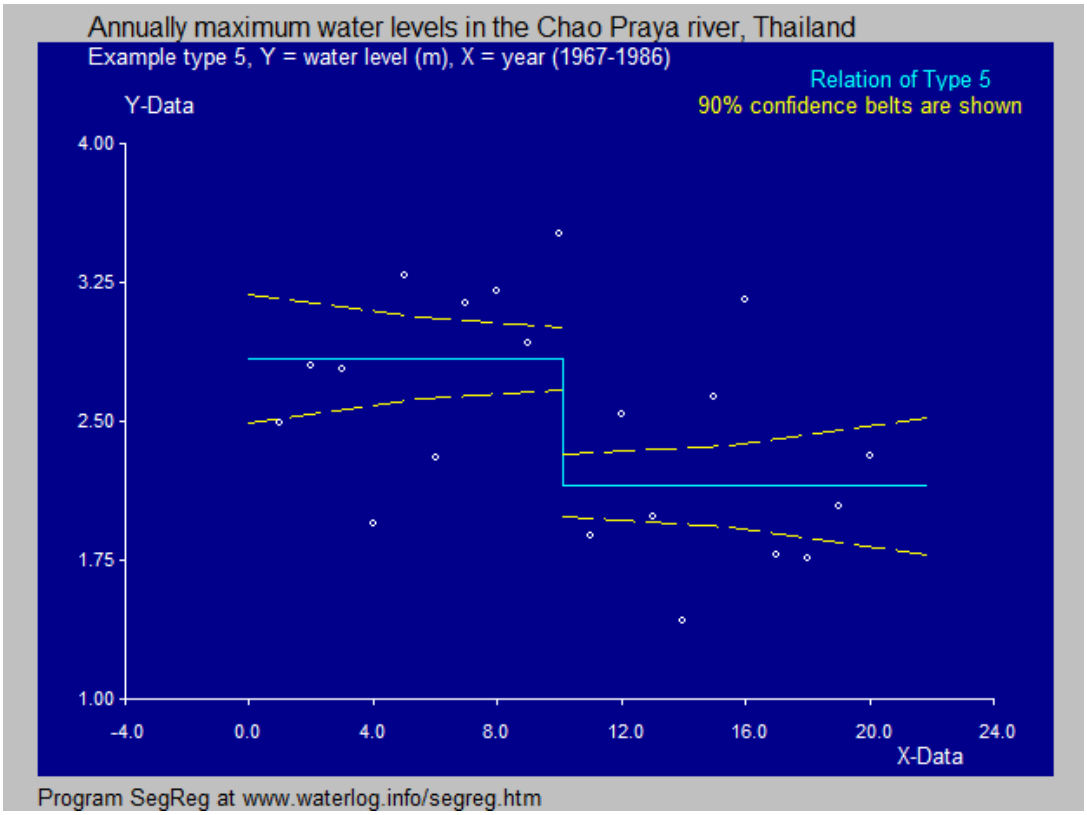
Notes

1. A major disadvantage of the F-test compared to the confidence-area test of BP, as done in SegReg, is that one obtains only a yes (BP is likely) or no (BP is doubtful) answer and one lacks an insight into the error-range to which the BP might be subjected even when it is significant.

2. $SSD_2 = SSD_6 - SSD_1$
   $SSD_3 = SSD_2 - SSD_4$
   $SSD_5 = SSD_6 - SSD_4$

Example of numerical ANOVA table

The following example of a numerical ANOVA table is taken from the data file of the SegReg program



Example of a segmented regression of Type 5 using the SegReg program

The general data are as shown in the next table

Variance Analysis, ANOVA table, Regression Type: 5 Sum[(Y-Av.Y)sq.] = 6.820 (total sum of squares of deviations) Total nr. of data = 20 Degrees of freedom = 19 (one degree is used for parameter Av.Y

Using the same colour for corresponding values, The ANOVA table looks as follows (there may be slight differences in some outcomes due to rounding off):

Sum of squares of deviations (SSD)	Degrees of freedom (df)	Variance (Var)	F-test	Probability or Significance or Reliability (%)
Explained by linear regression: **1.070**	**1** (the slope uses 1 more df)	**1.070** / **1** = **1.070**	F (**1**, **18**) = **1.070** / **0.319** = 3.350	91.6 % slightly over 90% just significant
Remaining unexplained: 6.820 – **1.070** = **5.750**	19 – 1 = **18**	**5.750** / **18** = **0.319**		
Extra explained by break point: **1.276**	1 (The 2nd Y average uses another df)	**1.276** / **1** = **1.276**	F (**1**, **17**) = **1.276** / **0.262** = 4.951	96.1 % more than 95% highly significant very reliable
Remaining unexplained 19.700 – 1.276 = **4.453**	18 -1= **17**	**4.453** / **17** = **0.262**		

Note

It is repeated that owing to the use of confidence intervals in SegReg, application of F-tests is not strictly necessary. For example, when SegReg finds that the introduction of a break-point gives no significant additional explanation, because the confidence interval of BP is too wide, it will not show a breakpoint. Also, when it is found that simple linear regression gives no significant explanation, because the confidence interval of the regression coefficient is too wide, it will not use the regression.

Reference

G.W. Snedecor and W.G. Cochran (1980), Statistical Methods, 7th edition, Chap. 17.4, p. 401-403. Iowa State University Press.

Chap. 17.4 deals with “Extension of the analysis of variance in multiple linear regression to each individual explanatory variable”.