Statistical significance of segmented linear regression with break-point.

using variance analysis and F-tests.

On website www.waterlog.info , R.J.Oosterbaan

Subjects:

Introduction

In SegReg 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). 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

Symbols used

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 $R = r^2$
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)
δ	residual after segmented linear regression with break-point, also called deviation from segmented linear regression
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 , 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.

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.

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: x_a , b_b , y_a , y_b , η_a , η_b , r_a , r_b .

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
1.	total variation *) (initial, without regression)	$\Sigma(y-\eta)^2$	n-1	$SSD_{\mathbf{1}}/(n-1)$
2.	explanation by simple linear regression without BP ^)	$r^2\,\Sigma(y-\eta)^2$	**1**	$SSD_{\mathbf{2}}/1 = SSD_{\mathbf{2}}$
3.	remaining unexplained after linear regression (deviations or residuals from linear regression model)	$(1-r^2)\,\Sigma(y-\eta)^2$	n-2	$SSD_{\mathbf{3}}/(n-2)$

*) 1 df is lost for use of the mean

^) 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: r_a and r_b , and also two means: 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	Var
1.	total variation *) (initial, without regression)	$\Sigma(y-\eta)^2$	n-1	$SSD_{\mathbf{1}}/(n-1)$
2.	explanation by simple linear regression without BP ^)	$r^2\,\Sigma(y-\eta)^2$	1	$SSD_{\mathbf{2}}/1 = SSD_{\mathbf{2}}$
2.	unexplained (residual) after linear regression (deviations from linear regression model)	$(1- r_a^2)\,\Sigma(y_a-\eta_a)^2 + (1- r_b^2)\,\Sigma(y_b-\eta_b)^2$	n-2	$SSD_{\mathbf{3}}/(n-2)$
4.	additional explanation by BP analysis compared to simple linear regression #)	$(R-r_{a2})\,\Sigma(y_a-\eta_a)^2 + (R-r_b^2)\,\Sigma(y_b-\eta_b)^2$	**1**	$SSD_{\mathbf{4}}/1 = SSD_{\mathbf{2}}$
5.	unexplained (residual) after BP analysis Type 5 (deviations from BP regression model)	$(1-R)\,\Sigma(y_a-\eta_a)^2 + (1-R)\,\Sigma(y_b-\eta_b)^2$	n-3	$SSD_{\mathbf{5}}/(n-3)$
6.	total explanation by segmented linear regression with BP &)	$R.\Sigma(y-\eta)^2$	2	$SSD_{\mathbf{6}}/2$

*) 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: r_a and r_b , and also two means: 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	Var
1.	total variation (initial, without regression)	*)	$\Sigma(y-\eta)^2$	n-1	$SSD_{\mathbf{1}}/(n-1)$
2.	explanation by simple linear regression without BP	^)	$r^2\,\Sigma(y-\eta)^2$	1	$SSD_{\mathbf{2}}/1 = SSD_{\mathbf{2}}$
3.	unexplained (residual) after linear regression (deviations from linear regression model)		$(1- r_a^2)\,\Sigma(y_a-\eta_a)^2 + (1- r_b^2)\,\Sigma(y_b-\eta_b)^2$	n-2	$SSD_{\mathbf{3}}/(n-2)$
4.	additional explanation by BP analysis compared to simple linear regression	#)	$(R-r_{a2})\,\Sigma(y_a-\eta_a)^2 + (R-r_b^2)\,\Sigma(y_b-\eta_b)^2$	**2**	$SSD_{\mathbf{4}}/2$
5.	unexplained (residual) after BP analysis Type 5 (deviations from BP regression model)		$(1-R)\,\Sigma(y_a-\eta_a)^2 + (1-R)\,\Sigma(y_b-\eta_b)^2$	n-4	$SSD_{\mathbf{5}}/(n-4)$
6.	total explanation by segmented linear regression with BP	&)	$R.\Sigma(y-\eta)^2$	3	$SSD_{\mathbf{6}}/3$

*) 1 df is lost for use of the mean
^) Another df is lost for use of the slope (regression coefficient)
#) In types 3 and 4, **two** more degrees of freedom are lost owing to the use of the second mean and the one slope
@) 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. (two sloping lines).
Two correlation coefficients are used: r_a and r_b , and also two means: 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	Var
1.	total variation (initial, without regression)	*)	$\Sigma(y-\eta)^2$	n-1	$SSD_{\mathbf{1}}/(n-1)$
2.	explanation by simple linear regression without BP	^)	$r^2\,\Sigma(y-\eta)^2$	1	$SSD_{\mathbf{2}}/1 = SSD_{\mathbf{2}}$
3.	unexplained (residual) after linear regression (deviations from linear regression model)		$(1- r_a^2)\,\Sigma(y_a-\eta_a)^2 + (1- r_b^2)\,\Sigma(y_b-\eta_b)^2$	n-2	$SSD_{\mathbf{3}}/(n-2)$
4.	additional explanation by BP analysis compared to simple linear regression	#)	$(R-r_{a2})\,\Sigma(y_a-\eta_a)^2 + (R-r_b^2)\,\Sigma(y_b-\eta_b)^2$	**3**	$SSD_{\mathbf{4}}$
5.	unexplained (residual) after BP analysis Type 5 (deviations from BP regression model)		$(1-R)\,\Sigma(y_a-\eta_a)^2 + (1-R)\,\Sigma(y_b-\eta_b)^2$	n-5	$SSD_{\mathbf{5}}/(n-5)$
6.	total explanation by segmented linear regression with BP	&)	$R.\Sigma(y-\eta)^2$	4	$SSD_{\mathbf{6}}/4$

*) 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
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.

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

Reference

G.W. Snedecor and W.G. Cochran (1980), Statistical Methods, 7th edition, Chap. 19.3, p. 401-403. Iowa State University Press.

Chap. 19.3 deals with the “second degree polynomial”, which I have replaced by “segmented regression with break-point” as the principles are the same.

Analysis of variance for data having several y-values for each x-value

Additional symbols

a	number of classes or groups (minimum 4)
i	index indicating a group number (i = 1, 2, 3, …a)
γ_i	mean value of y of group i
λ_i	value of y at x_i according to segmented linear regression
x_i	x-value of group i
j_i	number of y-values in group i

Table 5. ANOVA table for data having several y-values for each x-value (Type 3 and 4 only)

Nr.	Description	SSD	df	Var
1.	explanation by simple linear regression without BP	$r^2 \Sigma(y-\eta)^2$	1	$SSD_{\mathbf{1}}$ (not used)
2.	additional explanation by BP analysis over simple linear regression	$(R-r^2) \Sigma(y-\eta)^2$	1	$SSD_{\mathbf{2}}$
3.	explanation by segmented linear regression with BP **1** + **2**	$R.\Sigma(y-\eta)^2$	2	$SSD_{\mathbf{3}}/2$ (not used)
4.	deviations of group means from segmented regression	$\Sigma(\gamma_i-\lambda)^2$	a-3	$SSD_{\mathbf{4}}/(a-3)$
5.	deviations between groups **3** + **4**	$SSD_{\mathbf{3}}+SSD_{\mathbf{4}}$	a-1	$SSD_{\mathbf{5}}/(a-1)$
6.	deviations within groups	$\Sigma_i \Sigma_{j_i} (y-\gamma_i)$	n-a	$SSD_{\mathbf{6}}/(n-a)$
7.	unexplained after regression with BP (deviations from regression model) **5** + **6**	$(1-R) \Sigma(y-\eta)^2$	n-3	$SSD_{\mathbf{7}}/(n-3)$
8.	total (initial, without regression)	$\Sigma(y-\eta)^2$	n-1	$SSD_{\mathbf{8}}/(n-1)$ (not used)

The test-statistic F_0 = Var2/Var7 derived from Table 2 has the same meaning as the statistic F_0 = Var4/Var5 from Table 1.

The additional test-statistic is F_a = Var4/Var6, that can be compared to the value of F(a-3,n-a) in the F-tables. To conclude that the segmented regression is probably justified Fa needs to be the smaller one.

When the statistic F_a is significant, i.e. > F(a-3,n-a), and F_0 is significant (i.e. the contribution of BP is significant), the apparent non-linearity is probably not properly described by the present regression model.

Reference

G.W. Snedecor and W.G. Cochran (1980), Statistical Methods, 7th edition, Chap. 19.4, p. 401-403. Iowa State University Press.

Chap. 19.4 deals with the "Data having several Ys for each X".

Analysis of variance when two independent variables are present

When two independent variables (X and Z) are present, SegReg calculates the segmented linear regression with breakpoint (SLRB) of Y upon X and Y upon Z and it detects which of the two offers the highest degree of explanation. If this is true for X, it then performs an SLRB of the residuals YXr (after SLRB of Y upon X) on Z and the results are joined into one model. If, instead, Z offers the best fit after the first trial, the second SLRB is that of the residuals YZr (after SLRB of Y upon Z).

The sum of squares of deviations after a simple linear regression of residuals Yr (either YXr or YZr) on respectively Z or X can be found from

$$\Sigma(\varepsilon_1)^2 = [(\text{St.Dev.YXr}).(n-1)]^2$$

$$\Sigma(\varepsilon_2)^2 = [(\text{St.Dev.YZr}).(n-1)]^2$$

The values of the standard deviations of YXr or YZr are found in the SegReg output menu in the data category where BP=0 in the regression with residuals.

In the output file, SegReg also mentions in the same category the squares of the correlation coefficients ("corr.coeff.sq"), that we will call r_1^2 for a simple linear regression of YXr on Z.

The correlation coefficient calculated for the simple linear regression of Y upon X (as defined on page 1) or Y upon Z will be denoted by r_x and r_y respectively.

Similarly, the coefficient of explanation after SLRB of Y on X or Y on Z will be indicated by R_x and R_z respectively. Their values can be found in the output file of SegReg in the category "parameters for functions of type …" just after the segmented linear regression with breakpoint of Y on X or Y on Z.

For the first case (regression of Y on X followed by a regression of the reduals YXr on Z) The residuals left and right of the breakpoint (BP) are indicated respectively by ε_{1a}, ε_{1b}, ε_{2a}, and ε_{2b}. The correlation coefficients are denoted by r_x, and r_1.

Further, the output file gives data on the coefficient of explanation ("expl.coeff") for the SLRB of either YXr on Z or YZr on X in the category "parameters for functions of type ….". These values we will call respectively R_1 and R_2.

In the summary of the output file of Segreg, the overall coefficient explanation for the combined effect of the SLRB on X or Z and the residuals YZr and YXr respectively is given as “overall coefficient of explanation. This factor we will call R_t.

Now we can prepare an ANOVA table, depending on whether the first regression was done with the variable X or with the variable Z.

For the first case, assuming that both the regression of Y on X and YXr on Z give Type 5, we get:

Table 6. ANOVA table for SLRB of Y upon X and the residuals YXr on Z

Nr.	Description	SSD	df	Var
1.	explanation by simple linear regression of Y on X without BP	$r_x^2\,\Sigma(y-\eta)^2$	1	$SSD_{\mathbf{1}}$
2.	unexplained after linear regression of Y on X without BP (deviations from regression model)	$(1-r_x^2)\,\Sigma(y-\eta)^2$	n-2	$SSD_{\mathbf{2}}/(n-2)$
3.	additional explanation by BP analysis over simple linear regression of Y on X	$(R_x-r_{xa}^2)\,\Sigma(y_a-\eta_a)^2$ + $(R_x-r_{xb}^2)\,\Sigma(y_b-\eta_b)^2$	1	$SSD_{\mathbf{3}}$
4.	unexplained after SLRB analysis of Y on X (deviations from regression model)	$(1-R_x)\,\Sigma(y_a-\eta_a)^2$ + $(1-R_x)\,\Sigma(y_b-\eta_b)^2$	n-3	$SSD_{\mathbf{4}}/(n-3)$
5.	total explanation by segmented linear regression with BP of Y on X	$R_x.\Sigma(y_a-\eta_a)^2$ + $R_x.\Sigma(y_b-\eta_b)^2$	n-3	$SSD_{\mathbf{5}}/(n-3)$
6.	explanation by simple linear regression of YXr on Z without BP	$r_1^2\,\Sigma(\varepsilon_1)^2$	1	$SSD_{\mathbf{6}}$
7.	unexplained after linear regression of YXr on Z without BP	$(1-r_1^2)\,\Sigma(\varepsilon_1)^2$	n-4	$SSD_{\mathbf{7}}/(n-4)$
8.	additional explanation by BP analysis over simple linear regression of YXr on Z	$(R_1-r_{1a}^2)\,\Sigma(\varepsilon_{1a})^2$ + $(R_1-r_{1b}^2)\,\Sigma(\varepsilon_{1b})^2$	1	$SSD_{\mathbf{8}}$
9.	unexplained after SLRB analysis of YXr on Z (deviations from regression model)	$(1-R_1)\,\Sigma(\varepsilon_{1a})^2$ + $(1-R_1)\,\Sigma(\varepsilon_{1b})^2$	n-5	$SSD_{\mathbf{9}}/(n-5)$
10	total explanation by segmented linear regression with BP of YXr on Z	$R_2.\Sigma(\varepsilon_{1a})^2$ + $R_2.\Sigma(\varepsilon_{1b})^2$	2	$SSD_{\mathbf{10}}/2$

Now, the following test values of F can be prepared:

F1(1,n-2) = Var1/Var2, to see if the simple linear regression of Y on X is significant
F2(1,n-3) = Var3/Var4, to see if the SLRB of Y on X is significant
F3(1,n-4) = Var6/Var7, to see if the simple linear regression of YXr on Z is significant
F4(1,n-5) = Var8/Var9, to see if the SLRB of YXr on Z is significant.

Note 1
For the second case (first a regression of Y upon Z followed by a regression of residuals YZr on X, the procedure, mutatis mutandis, is similar as shown in Table 6

Note 2
Table 6 has shown an example when both segmented regressions (Y on X or Z followed by Y-residuals on Z or X) are of Type 5. When other types appear, the lost number of degrees of freedom needs to be adjusted in accordance to the procedures shown in Tables 2, 3 and 4.

Note 3
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”.