Segmented Regression

Steven Lillywhite

June 20, 2021

Abstract

We review univariate regression models where the functional form is piecewise linear and continuous. We do not strive for mathematical completeness, but focus rather on explaining some of the formulas behind the implementation in the Python project `segreg`.

Contents

1 Introduction

Segmented regression models may be used to study changes in regime, or simply as more flexible fits to non-linear data. There are exact algorithms for parameter estimation that reduce to multiple ordinary least squares regressions and function evaluations. So parameter estimation is fast and exact, in contrast to the non-linear optimization routines that can arise with other types of non-linear models. For an introduction to the subject, we mention [5], [10], [7] and the references contained therein.

In this document, we shall only describe segmented regression for univariate data. We note that all of the information we present is well known. We are only providing a review and explaining details of the algorithms.

When modelling data with segmented regression models, a common problem is determining the number of breakpoints. For this we prefer the Bayesian Information Criterion (BIC), for which we provide a brief overview in Sections 6 and 7.

1.1 Regression Model

The models we shall consider have the general form

$$y(x) = f(x) + \varepsilon(x) \tag{1.1}$$

where x and y are univariate. We assume the residual ε has mean zero and variance σ^2. We shall further assume we have an N observations from an i.i.d. sample $\{(y_i, x_i)\}_{i=1}^N$.

1.2 Segmented Regression Model

We only consider segmented regression models for which $f(x)$ is piecewise linear. The function $f(x)$ will have the general form

$$f(x) = \begin{cases} v_1 + m_1(x - u_1) & x \le u_1 \\ v_2 + m_2(x - u_2) & u_1 \le x \le u_2 \\ v_3 + m_3(x - u_3) & u_2 \le x \le u_3 \\ \quad \vdots & \\ v_k + m_k(x - u_k) & u_{k-1} \le x \le u_k \\ v_k + m_{k+1}(x - u_k) & u_k \le x \end{cases} \tag{1.2}$$

We call the points where the functional form changes *breakpoints*. So the function (1.2) has k breakpoints $\{u_1, u_2, \ldots, u_k\}$. The way we have parameterized the function, the breakpoints in the x-y plane are $\{(u_1, v_1), (u_2, v_2), \ldots, (u_k, v_k)\}$

We will only consider continuous functions of this form. This implies further restrictions on the parameters which we will treat below.

The estimation formulas below can all be expressed using matrix operations. However, the sampling distributions of segmented regression parameters can often be highly skewed, which leads us to employ bootstrap techniques for calculation of confidence intervals and other statistical tests. Such tests involve many calculations of the parameter estimates via repeated simulations. As such, we wish to make the calculations as fast as possible. This has motivated us to formulate the parameter estimates without matrix operations. The resulting formulas are much faster to compute, albeit rather cumbersome.

1.3 Notation

For summations, an unadorned summation symbol $\sum$ will imply summation over all possible indices, ie: $\sum \stackrel{\text{def}}{=} \sum_{i=1}^{N}$

We shall use integer labels under summation signs to indicate that the sums are over sets as defined in the following table.

Type	Label	Indices
One Breakpoint	1	$\{x_i \vert x_i \leq u_1\}$
	2	$\{x_i \vert u_1 < x_i\}$
Two Breakpoints	1	$\{x_i \vert x_i \leq u_1\}$
	2	$\{x_i \vert u_1 < x_i \leq u_2\}$
	3	$\{x_i \vert u_2 < x_i\}$

We shall denote

- the cardinality of each set as $N_k \stackrel{\text{def}}{=} \sum_k 1$
- the indicator function by
$$\mathbb{1}_A(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$$
- estimates of parameters with hat notation, eg: $\widehat{\beta}$

We shall abbreviate ordinary least squares as OLS, residual sum of squares as RSS, maximum likelihood estimate as MLE, and probability distribution function as pdf.

2 No Breakpoint

The simplest case of segmented regression is when there is no breakpoint. This case is the standard ordinary least squares. We may express the model as

$$Y = X\beta + \boldsymbol{\varepsilon}$$

where

$$Y_{N\times 1} \stackrel{\text{def}}{=} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \qquad X_{N\times 2} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{pmatrix} \qquad \beta_{2\times 1} \stackrel{\text{def}}{=} \begin{pmatrix} v \\ m \end{pmatrix} \qquad \boldsymbol{\varepsilon}_{N\times 1} \stackrel{\text{def}}{=} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

We label the components of β in this way in order to be consistent with parameter labeling we use for segmented regression. The estimate for β is then

$$\widehat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$$

It is easy to see that the matrix operations will involve some redundancies. With a view to speeding calculation, especially for the segmented regressions with breakpoints, we will solve the regression in the following way.

Lemma 1
Consider the ordinary least squares model

$$y = v + mx + \varepsilon$$

The regression estimates for the parameters v and m are

$$\widehat{m} = \frac{N\sum y_i x_i - \sum y_i \sum x_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\widehat{v} = \frac{1}{N}\sum y_i - \frac{\widehat{m}}{N}\sum x_i$$

Proof. The regression parameter estimates are obtained by minimizing the RSS which we write as

$$RSS \stackrel{\text{def}}{=} G(v,m) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \left[y_i - (v + mx_i)\right]^2$$

So we wish to solve the following optimization problem.

$$\underset{v,m}{\operatorname{argmin}}\ G(v,m)$$

One may show that $G(v,m)$ is convex, implying that it has a unique global minimum. Solving for critical points of the function we have

$$\frac{\partial G}{\partial v} = -2\sum_{i=1}^{N} \left[y_i - (v + mx_i)\right] = 0$$
$$\frac{\partial G}{\partial m} = -2\sum_{i=1}^{N} \left[y_i - (v + mx_i)\right] x_i = 0$$

which yields

$$v = \frac{1}{N}\sum \left[y_i - mx_i\right]$$
$$m = \frac{\sum \left[y_i - v\right] x_i}{\sum_{i=1}^{N} x_i^2}$$

Substituting the expression for v into the expression for m, we obtain estimators (denoted with hat)

$$\widehat{m} \stackrel{\text{def}}{=} \frac{N\sum y_i x_i - \sum y_i \sum x_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{2.1}$$
$$\widehat{v} \stackrel{\text{def}}{=} \frac{1}{N}\sum y_i - \frac{\widehat{m}}{N}\sum x_i$$

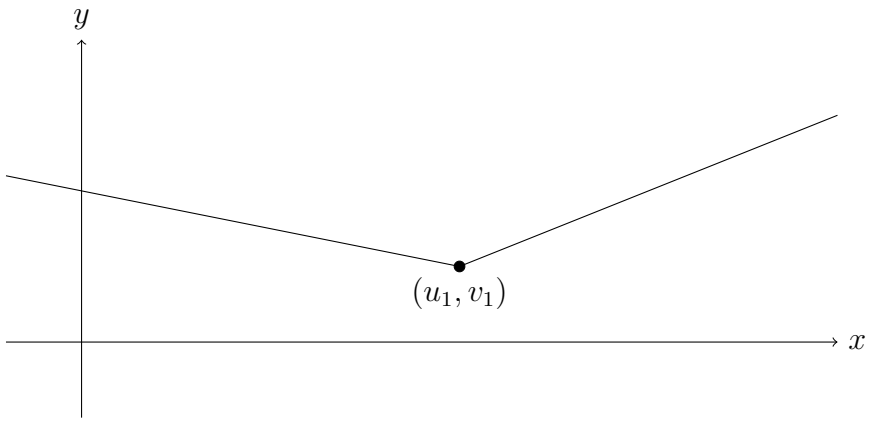
□

3 One Breakpoint

With only one breakpoint, the model takes the form

$$f(x) = \begin{cases} v_1 + m_1(x - u_1) & x \leq u_1 \\ v_1 + m_2(x - u_1) & u_1 \leq x \end{cases} \tag{3.1}$$

Here we have parametrized two half-lines with slopes m_1, m_2 which meet at a common vertex (u_1, v_1).



We will estimate parameters by means of regression. To this end, we define the residual sum of squares function as

$$\begin{aligned} G(u_1, v_1, m_1, m_2) &\stackrel{\text{def}}{=} \sum (y_i - f(x_i))^2 \\ &= \sum_1 [y_i - v_1 - m_1(x_i - u_1)]^2 + \sum_2 [y_i - v_1 - m_2(x_i - u_1)]^2 \end{aligned} \tag{3.2}$$

3.1 Estimate for Fixed Breakpoint

When solving for parameters which minimize G, if we fix u_1, the remaining parameters that minimize G are easily determined. Suppose we fix a value for u_1 such that $x_i < u_1 < x_{i+1}$. Then we may write $f(x)$ in the form

$$f(x) = v_1 + m_1(x - u_1)\mathbb{1}_{x \leq u_1} + m_2(x - u_1)\mathbb{1}_{u_1 < x} \tag{3.3}$$

where v_1, m_1, m_2 are the unknown parameters and $\mathbb{1}$ is the indicator function defined as

$$\mathbb{1}_A(t) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{else} \end{cases}$$

3.1.1 Regression Formula

We may apply OLS to equation (3.3) to determine the parameters v_1, m_1, m_2 minimizing (3.2) for this fixed value of u_1. In matrix form, it looks like

$$Y = X\beta + \boldsymbol{\varepsilon}$$

$$X_{N\times 3} \stackrel{\text{def}}{=} \left(\begin{array}{c|c|c} 1 & x_1 - u_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_i - u_1 & 0 \\ \hline 1 & 0 & x_{i+1} - u_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_N - u_1 \end{array}\right) \qquad \beta_{3\times 1} \stackrel{\text{def}}{=} \begin{pmatrix} v_1 \\ m_1 \\ m_2 \end{pmatrix}$$

3.1.2 Brute Force Calculation

As mentioned in Section 1.2, the typical regression matrix solution will involve many redundant calculations, and can be faster computationally to solve directly for the parameters by brute force calculation. The formulas are shown in Section 5.2.

3.2 Estimate for Unknown Breakpoint

Let us denote the function of u_1 which computes the RSS with u_1 considered fixed, as $g(u_1)$. That is

$$g(u_1) \stackrel{\text{def}}{=} \operatorname*{argmin}_{v_1,m_1,m_2} G(u_1, v_1, m_1, m_2) \tag{3.4}$$

Minimizing the function $G(u_1, v_1, m_1, m_2)$ globally over all the parameters $\{u_1, v_1, m_1, m_2\}$ is the same as finding u_1 which minimizes $g(u_1)$. However, there are some difficulties in using $g(u_1)$ directly to find the minimum. Although $g(u_1)$ is continuous, it is unfortunately not differentiable at the data points $u_1 = x_i$. Furthermore, it can possess multiple local minima, even for moderate sample sizes. Nonetheless, it is possible to give a closed-form algorithm which solves for the global minimum of G.

3.2.1 Hudson's Algorithm

We present a closed-form algorithm due to Hudson which solves for the minimum of the RSS, G. This was first described in [2], but see also [10] and [7] for more description and generalizations of the method.

Let us order the data $x_1 < x_2 < \cdots < x_N$. It is easy to see that for $x_1 < u_1 < x_2$ the minimizer of G is given by taking the the right-hand-side line to be that determined by OLS for the data $\{(x_2, y_2), \ldots, (x_N, y_N)\}$, and taking the left-hand-side line to be the straight line

from (x_1, y_1) to the right-hand-side line. A similar result holds for $x_{N-1} < u_1 < x_N$. So $g(u_1)$ is constant on the intervals $[x_1, x_2]$ and $[x_{N-1}, x_N]$.

Next we consider the intervals $x_i < u_1 < x_{i+1}$ for $i = 2, 3, \ldots, N-2$. On such an interval, the function G is smooth. (Simple calculation shows that $\frac{\partial G}{\partial u_1}$ is discontinous at $u_1 = x_i$). To check for critical points of the function G on the interval $x_j < u_1 < x_{j+1}$, we solve

$$\frac{\partial G}{\partial u_1} = \frac{\partial G}{\partial v_1} = \frac{\partial G}{\partial m_1} = \frac{\partial G}{\partial m_2} = 0 \tag{3.5}$$

$\frac{\partial G}{\partial u_1} = 0$ implies

$$m_1 \sum_1 [y_i - v_1 - m_1(x_i - u_1)] + m_2 \sum_2 [y_i - v_1 - m_2(x_i - u_1)] = 0 \tag{3.6}$$

$\frac{\partial G}{\partial v_1} = 0$ implies

$$\sum_1 [y_i - v_1 - m_1(x_i - u_1)] + \sum_2 [y_i - v_1 - m_2(x_i - u_1)] = 0 \tag{3.7}$$

$\frac{\partial G}{\partial m_1} = 0$ implies

$$\sum_1 [y_i - v_1 - m_1(x_i - u_1)] (x_i - u_1) = 0$$

$\frac{\partial G}{\partial m_2} = 0$ implies

$$\sum_2 [y_i - v_1 - m_2(x_i - u_1)] (x_i - u_1) = 0$$

Now, (3.6) and (3.7) together imply

$$\left(1 - \frac{m_2}{m_1}\right) \sum_2 [y_i - v_1 - m_2(x_i - u_1)] = 0$$

There are two cases to consider. First assume $m_1 \neq m_2$. Then

$$\sum_2 [y_i - v_1 - m_2(x_i - u_1)] = 0$$

and (3.7) implies

$$\sum_1 [y_i - v_1 - m_1(x_i - u_1)] = 0$$

So the system (3.5) becomes

$$\begin{cases} \sum_1 [y_i - v_1 - m_1(x_i - u_1)] = 0 \\ \sum_1 [y_i - v_1 - m_1(x_i - u_1)] (x_i - u_1) = 0 \end{cases} \tag{3.8}$$

$$\begin{cases} \sum\limits_{2} \left[y_i - v_1 - m_2(x_i - u_1) \right] = 0 \\ \sum\limits_{2} \left[y_i - v_1 - m_2(x_i - u_1) \right] (x_i - u_1) = 0 \end{cases} \tag{3.9}$$

But (3.8) are just the equations which solve for OLS for the data set $\{(x_1, y_1), \ldots (x_j, y_j)\}$ and (3.9) are the equations which solve for OLS for the data set $\{(x_{j+1}, y_{j+1}), \ldots (x_N, y_N)\}$. We can easily see this by parameterizing the lines using the point-slope formula with the point being (u_1, v_1) and the slopes m_1 and m_2 respectively. Setting the derivatives with respect to the parameters equal to zero gives these equations (the *normal equations*).

We compute the closed-form OLS solutions for these two data sets. The slopes of the resulting lines and the intersection of the two lines gives the critical points of G. If the u_1 coordinate of the intersection of the two OLS lines lies in the interval (x_j, x_{j+1}), then we have found a critical point for G on this interval. Otherwise there are no critical points for G on this interval.

If $m_1 = m_2$, we are in the case of a line fit (ie: OLS) to the entire data set. Then every point value for u_1 on the interval (x_j, x_{j+1}) gives a minimum for G, and consequently the function $g(u_1)$ is constant on the interval.

3.2.2 Algorithm For One-Breakpoint Segmented Regression

We thus have the following algorithm to find the argmin of G (3.2) by comparing the values of g at all the critical points.

1. For each interval $x_i < u_1 < x_{i+1}$ for $i = 2, 3, \ldots, N-2$ solve OLS for the left-hand data set and the right-hand data set. If the intersection of the two lines lies in the interval (x_i, x_{i+1}), then record the resulting value for g (3.4).

   More precisely, solve OLS on the left-hand data set $\{x_1, \ldots, x_i\}$, obtaining a line fit with slope $\widehat{m}_1$. Similarly solve OLS and on the right-hand data set $\{x_{i+1}, \ldots, x_N\}$, obtaining a line fit with slope $\widehat{m}_2$. Find the intersection of the lines $(\widehat{u}_1, \widehat{v}_1)$. If $x_i < \widehat{u}_1 < x_{i+1}$, then we record $g(\widehat{u}_1) = G(\widehat{u}_1, \widehat{v}_1, \widehat{m}_1, \widehat{m}_2) = \text{RSS}$.

2. Record the values of g at the data points: $g(x_1), g(x_2), \ldots, g(x_N)$.

3. Find the value of u_1 which gives the minimum value for g among the recorded values in the previous steps. This is the solution.

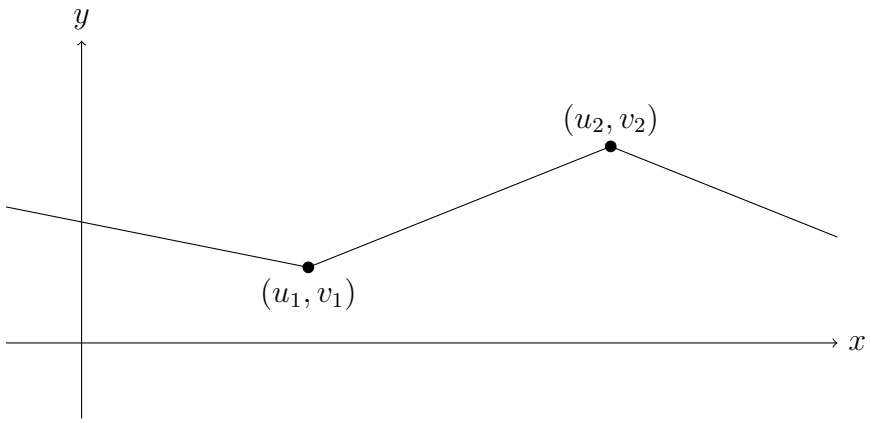
We note that in practice, we add the restriction that there must be a minimum number of distinct data points in each linear segment of the model. This avoids pathological fits and overfitting that we would not wish to consider valid.

4 Two Breakpoints

With two breakpoints, the model takes the form

$$f(x) = \begin{cases} v_1 + m_1(x - u_1) & x \leq u_1 \\ v_1 + \left(\dfrac{v_2 - v_1}{u_2 - u_1}\right)(x - u_1) & u_1 \leq x \leq u_2 \\ v_2 + m_2(x - u_2) & u_2 \leq x \end{cases} \tag{4.1}$$

Here there are three line segments. Continuity of $f(x)$ allows us to parameterize the slope of the middle segment in terms of the two endpoints (u_1, v_1) and (u_2, v_2).



There are thus six free parameters, and the residual sum of squares function is

$$\begin{aligned} G(u_1, v_1, u_2, v_2, m_1, m_2) &\stackrel{\text{def}}{=} \sum (y_i - f(x_i))^2 \\ &= \sum_1 [y_i - v_1 - m_1(x_i - u_1)]^2 + \\ &\quad \sum_2 \left[y_i - v_1 - \left(\frac{v_2 - v_1}{u_2 - u_1}\right)(x_i - u_1)\right]^2 + \\ &\quad \sum_3 [y_i - v_2 - m_2(x_i - u_2)]^2 \end{aligned}$$

4.1 Estimate for Fixed Breakpoints

Similarly to the case for a single breakpoint, when solving for parameters which minimize G, it turns out that we may concentrate out v_1, v_2, m_1, m_2. That is, conditional on a values for u_1 and u_2, the remaining parameters are all determined.

4.1.1 Regression Formula

Suppose we fix u_1, u_2 such that

$$x_i < u_1 < x_{i+1} < x_j < u_2 < x_{j+1}$$

Then we may express the model (1.1) for the function (4.1) by writing the classical regression formula

$$Y = X\beta + \boldsymbol{\varepsilon}$$

with $Y_i = f(x_i)$ and

$$X_{N\times 4} \stackrel{\text{def}}{=} \left(\begin{array}{c|c|c|c} 1 & 0 & x_1 - u_1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_i - u_1 & 0 \\ \hline 1 - \frac{x_{i+1}-u_1}{u_2-u_1} & \frac{x_{i+1}-u_1}{u_2-u_1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 - \frac{x_j-u_1}{u_2-u_1} & \frac{x_j-u_1}{u_2-u_1} & 0 & 0 \\ \hline 0 & 1 & 0 & x_{j+1} - u_2 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_N - u_2 \end{array}\right) \qquad \beta_{4\times 1} \stackrel{\text{def}}{=} \begin{pmatrix} v_1 \\ v_2 \\ m_1 \\ m_2 \end{pmatrix}$$

4.1.2 Brute Force Calculation

However, as mentioned in Section 1.2, the typical regression matrix solution will involve many redundant calculations, and it can be faster to solve directly for the parameters by brute force calculation. The resulting formulas are shown in Section 5.3.

4.2 Estimate for Unknown Breakpoints

We fit the model again by minimizing the RSS. This time Hudson's algorithm involves a two-dimensional grid, but each step ultimately reduces to OLS regressions and function evaluations, as before.

5 Formulas

We collect the brute-force formulas here to provide an easy reference for coding them.

5.1 No Breakpoint

FORMULA

$$y = v + mx + \varepsilon$$

PRECOMPUTE

$$N \qquad \sum x_i \qquad \sum x_i^2 \qquad \sum y_i \qquad \sum y_i^2 \qquad \sum y_i x_i$$

DEFINE

$$A \stackrel{\text{def}}{=} \sum x_i \qquad B \stackrel{\text{def}}{=} \sum x_i^2 \qquad C \stackrel{\text{def}}{=} \sum y_i \qquad D \stackrel{\text{def}}{=} \sum y_i^2 \qquad E \stackrel{\text{def}}{=} \sum y_i x_i$$

ESTIMATES

$$\widehat{m} = \frac{NE - AC}{NB - A^2}$$
$$\widehat{v} = \frac{C}{N} - \widehat{m}\frac{A}{N}$$

RSS

$$\sum [y_i - \widehat{v} - \widehat{m}x_i]^2 = D - 2\widehat{v}C - 2\widehat{m}E + \widehat{v}^2 N + 2\widehat{v}\widehat{m}A + \widehat{m}^2 B$$

5.2 One Breakpoint

FORMULA

$$f(x) = \begin{cases} v_1 + m_1(x - u_1) & x \le u_1 \\ v_1 + m_2(x - u_1) & u_1 \le x \end{cases}$$

PRECOMPUTE
(for $k = 1, 2$)

$$N_k \qquad \sum_k x_i \qquad \sum_k x_i^2 \qquad \sum_k y_i \qquad \sum_k y_i^2 \qquad \sum_k y_i x_i$$

DEFINE
(for $k = 1, 2$)

$$\begin{aligned} A &\stackrel{\text{def}}{=} \sum y_i \\ B_k &\stackrel{\text{def}}{=} \sum_k y_i(x_i - u_1) = \sum_k y_i x_i - u_1 \sum_k y_i \\ C_k &\stackrel{\text{def}}{=} \sum_k (x_i - u_1) = \sum_k x_i - u_1 N_k \\ D_k &\stackrel{\text{def}}{=} \sum_k (x_i - u_1)^2 = \sum_k x_i^2 - 2u_1 \sum_k x_i + u_1^2 N_k \\ E &\stackrel{\text{def}}{=} \sum y_i^2 \end{aligned}$$

ESTIMATES FOR FIXED u_1

$$\widehat{v}_1 = \frac{A - \dfrac{B_1 C_1}{D_1} - \dfrac{B_2 C_2}{D_2}}{N - \dfrac{C_1^2}{D_1} - \dfrac{C_2^2}{D_2}} \qquad \widehat{m}_1 = \frac{B_1 - v_1 C_1}{D_1} \qquad \widehat{m}_2 = \frac{B_2 - v_1 C_2}{D_2} \tag{5.1}$$

RSS

$$\sum_1 \left[y_i - \widehat{v}_1 - \widehat{m}_1(x_i - u_1)\right]^2 + \sum_2 \left[y_i - \widehat{v}_1 - \widehat{m}_2(x_i - u_1)\right]^2 =$$
$$E - 2\widehat{v}_1 A + \widehat{v}_1^2 N - 2\widehat{m}_1 B_1 - 2\widehat{m}_2 B_2 + 2\widehat{v}_1\widehat{m}_1 C_1 + 2\widehat{v}_1\widehat{m}_2 C_2 + \widehat{m}_1^2 D_1 + \widehat{m}_2^2 D_2$$

5.3 Two Breakpoints

FORMULA

$$f(x) = \begin{cases} v_1 + m_1(x - u_1) & x \le u_1 \\ v_1 + \left(\dfrac{v_2 - v_1}{u_2 - u_1}\right)(x - u_1) & u_1 \le x \le u_2 \\ v_2 + m_2(x - u_2) & u_2 \le x \end{cases}$$

PRECOMPUTE
(for $k = 1, 2, 3$)

$$N_k \qquad \sum_k x_i \qquad \sum_k x_i^2 \qquad \sum_k y_i \qquad \sum_k y_i^2 \qquad \sum_k y_i x_i$$

DEFINE
(for $k = 1, 2, 3$ and $r = 1, 2$)

$$A_k \stackrel{\text{def}}{=} \sum_k y_i$$
$$B_{k,r} \stackrel{\text{def}}{=} \sum_k y_i(x_i - u_r) = \sum_k y_i x_i - u_r \sum_k y_i$$
$$C_{k,r} \stackrel{\text{def}}{=} \sum_k (x_i - u_r) = \sum_k x_i - u_r N_k$$
$$D_{k,r} \stackrel{\text{def}}{=} \sum_k (x_i - u_r)^2 = \sum_k x_i^2 - 2u_r \sum_k x_i + u_1^2 N_k$$
$$E \stackrel{\text{def}}{=} \sum y_i^2$$
$$F_k \stackrel{\text{def}}{=} \sum_k (x_i - u_1)(x_i - u_2) = \sum_k x_i^2 - (u_1 + u_2) \sum_k x_i + u_1 u_2 N_k$$

ESTIMATES FOR FIXED u_1, u_2

Further define

$$
\begin{aligned}
a &\stackrel{\text{def}}{=} -N_1 + \frac{C_{1,1}^2}{D_{1,1}} - \frac{D_{2,2}}{(u_2-u_1)^2} \\
b &\stackrel{\text{def}}{=} \frac{F_2}{(u_2-u_1)^2} \\
c &\stackrel{\text{def}}{=} \frac{F_2}{(u_2-u_1)^2} \\
d &\stackrel{\text{def}}{=} -N_3 + \frac{C_{3,2}^2}{D_{3,2}} - \frac{D_{2,1}}{(u_2-u_1)^2} \\
e &\stackrel{\text{def}}{=} -A_1 + \frac{B_{1,1}C_{1,1}}{D_{1,1}} + \frac{B_{2,2}}{(u_2-u_1)} \\
f &\stackrel{\text{def}}{=} -A_3 + \frac{B_{3,2}C_{3,2}}{D_{3,2}} - \frac{B_{2,1}}{(u_2-u_1)}
\end{aligned}
$$

Then $\widehat{v}_1$ and $\widehat{v}_2$ are determined by solving the linear system

$$
\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \widehat{v}_1 \\ \widehat{v}_2 \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix}
$$

and

$$
\widehat{m}_1 = \frac{B_{1,1} - \widehat{v}_1 C_{1,1}}{D_{1,1}} \qquad \widehat{m}_2 = \frac{B_{3,2} - \widehat{v}_2 C_{3,2}}{D_{3,2}}
$$

RSS

Let $\widehat{m} \stackrel{\text{def}}{=} \dfrac{\widehat{v}_2 - \widehat{v}_1}{u_2 - u_1}$

$$
\begin{aligned}
&\sum_1 [y_i - \widehat{v}_1 - \widehat{m}_1(x_i - u_1)]^2 + \sum_2 [y_i - \widehat{v}_1 - \widehat{m}(x_i - u_1)]^2 \\
&\qquad + \sum_3 [y_i - \widehat{v}_2 - \widehat{m}_2(x_i - u_2)]^2 = \\
&E - 2\widehat{v}_1(A_1 + A_2) - 2\widehat{v}_2 A_3 - 2\widehat{m}_1 B_{1,1} - 2\widehat{m} B_{2,1} - 2\widehat{m}_2 B_{3,2} \\
&+ \widehat{v}_1^2(N_1 + N_2) + \widehat{v}_2^2 N_3 + 2\widehat{v}_1(\widehat{m}_1 C_{1,1} + \widehat{m} C_{2,1}) + 2\widehat{v}_2 \widehat{m}_2 C_{3,2} \\
&\widehat{m}_1^2 D_{1,1} + \widehat{m}^2 D_{2,1} + \widehat{m}_2^2 D_{3,2}
\end{aligned}
$$

6 Log Likelihood

For a regression model defined as in Section 1.1, let us suppose that f depends on a finite set of fixed model parameters we collect in a vector θ. We shall use the notation $f(x;\theta)$ to denote the dependence of the definition of the function on the parameters θ. Let us denote the pdf of ε by $F(z)$. Then, assuming for the moment that the residuals are normally distributed, the likelihood function for the set of (iid) observations is

$$\prod_{i=1}^{N} F(z_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z_i^2}{2\sigma^2}}$$

Then the loglikelihood evaluated at the data, $l(\theta)$, is given by

$$\begin{aligned} l(\theta) &\stackrel{\text{def}}{=} \sum \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - f(x_i;\theta))^2}{2\sigma^2} \right] \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum (y_i - f(x_i;\theta))^2 \end{aligned} \tag{6.1}$$

Considering the variance σ^2 as a variable, we concentrate it out of the loglikelihood as follows.

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (y_i - f(x_i;\theta))^2$$

Setting this expression to zero, we obtain

$$\sigma^2 = \frac{1}{N}\sum (y_i - f(x_i;\theta))^2$$

Substituting this back into the loglikelihood (6.1) gives the following expression for the value of the loglikelihood at the MLE

$$\begin{aligned} l(\widehat{\theta}) &= -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log\left(N\right) - \frac{N}{2} - \frac{N}{2}\log\left(\sum (y_i - f(x_i;\widehat{\theta}))^2\right) \\ &= -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log\left(N\right) - \frac{N}{2} - \frac{N}{2}\log\left(RSS\right) \end{aligned} \tag{6.2}$$

where we denote the MLE as $\widehat{\theta}$. We see that the loglikehood value at the MLE only depends on N and the RSS.

7 Review of Bayesian Information Criterion

We will give a brief overview of several versions of the Bayes Information Criterion (BIC). For more information, one may consult many textbooks and papers, including [4], [9], and references therein.

We begin by considering a set of r models

$$\{\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_r\}$$

with prior probabilities

$$p(\mathcal{M}_1), \ldots, p(\mathcal{M}_r)$$

We assume one of the models is the true model and

$$\sum_{i=1}^{r} p(\mathcal{M}_i) = 1$$

We also assume each model $\mathcal{M}_j$ is associated with a parameter vector θ_j, and let k_j denote the number of parameters in θ_j.

By Bayes' Theorem, the posterior probability that $\mathcal{M}_j$ is the true model, given the data D, is

$$p(\mathcal{M}_j|D) = \frac{p(D|\mathcal{M}_j)p(\mathcal{M}_j)}{\sum_{i=1}^{r} p(D|\mathcal{M}_i)p(\mathcal{M}_i)} \tag{7.1}$$

The most common situation is to assume that all models are equally likely a priori. Then we have

$$p(\mathcal{M}_j|D) = \frac{p(D|\mathcal{M}_j)}{\sum_{i=1}^{r} p(D|\mathcal{M}_i)} \tag{7.2}$$

Since the denominator in (7.2) is the same for each model, it follows that the model with the highest probability of being true is the one for which $p(D|\mathcal{M}_j)$ is largest.

We may compute

$$p(D|\mathcal{M}_j) = \int p(D|\theta_j, \mathcal{M}_j)p(\theta_j|\mathcal{M}_j)d\theta_j$$

The following approximation gives rise to the traditional BIC.

$$\log p(D|\mathcal{M}_j) \approx \log p(D|\widehat{\theta_j}, \mathcal{M}_j) - \frac{k_j}{2}\log N$$

where $\widehat{\theta_j}$ is the maximum likelihood estimate of θ_j and N is the size of the data. Using the notation for the log of the likelihood function, $l(\widehat{\theta_j}) \stackrel{\text{def}}{=} \log p(D|\widehat{\theta_j}, \mathcal{M}_j)$, we write

$$\log p(D|\mathcal{M}_j) \approx l(\widehat{\theta_j}) - \frac{k_j}{2}\log N$$

The traditional BIC is defined as -2 times this expression, namely

$$BIC_j \stackrel{\text{def}}{=} -2l(\widehat{\theta_j}) + k_j \log N \tag{7.3}$$

Then *minimizing* the BIC over the set of models is the same as maximizing $\log p(D|\mathcal{M}_j)$, hence $p(D|\mathcal{M}_j)$, hence $p(\mathcal{M}_j|D)$, and so selects the model with the highest posterior probability.

Using the BIC approximation with (7.2), we may compute the model probabilities as

$$p(\mathcal{M}_j|D) = \frac{p(D|\mathcal{M}_j)}{\sum_{i=1}^{r} p(D|\mathcal{M}_i)} = \frac{e^{l(\widehat{\theta_j}) - \frac{k_j}{2}\log N}}{\sum_{i=1}^{r} e^{l(\widehat{\theta_i}) - \frac{k_i}{2}\log N}} \tag{7.4}$$

Remark 1
The better the fit $l(\widehat{\theta}_j)$, the lesser the BIC. However, this is weighed against the term $k_i \log N$, which we view as a penalty term involving the number of parameters in the model.

7.1 Bayesian Model Averaging

For some quantity of interest, Δ, such as a predicted future observation, we have

$$p(\Delta|D) = \sum_{i=1}^{r} p(\Delta|D, \mathcal{M}_i) p(\mathcal{M}_i|D)$$

So the distribution $p(\Delta|D)$ is a mixture distribution of each model $p(\Delta|D, \mathcal{M}_i)$ with weights $p(\mathcal{M}_i|D)$. In like manner we have the expected values

$$E[\Delta|D] = \sum_{i=1}^{r} E[\Delta|D, \mathcal{M}_i] p(\mathcal{M}_i|D)$$

7.2 BIC Properties

Notice that if we add a constant to all the exponent terms in (7.4), the model probabilities are unchanged. Equivalently, adding a constant term to the BIC does not affect the model selection.

For a set of regression models $y = f_i(x) + \epsilon_i$, consider the loglikelihood of the j^{th} model

$$l(\widehat{\theta}_j) = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log(N) - \frac{N}{2} - \frac{N}{2}\log(RSS_j)$$

(where we have dropped the dependency of RSS on the estimated parameter vector $\widehat{\theta}$). Since the term

$$-\frac{N}{2}\log(2\pi) + \frac{N}{2}\log(N) - \frac{N}{2}$$

is the same for all models, we can subtract it from each model's BIC and write (7.3) as

$$BIC_j \stackrel{\text{def}}{=} N\log(RSS_j) + k_j \log N \tag{7.5}$$

and simlarly write (7.4) as

$$p(\mathcal{M}_j|D) = \frac{e^{-\frac{N}{2}\log(RSS_j) - \frac{k_j}{2}\log N}}{\sum_{i=1}^{r} e^{-\frac{N}{2}\log(RSS_i) - \frac{k_i}{2}\log N}}$$

We can re-express this in various ways. For example

$$p(\mathcal{M}_1|D) = \frac{e^{-\frac{N}{2}\log(RSS_1)-\frac{k_1}{2}\log N}}{\sum_{i=1}^{r} e^{-\frac{N}{2}\log(RSS_i)-\frac{k_i}{2}\log N}} \tag{7.6}$$

$$= \frac{1}{1+\left(\frac{RSS_1}{RSS_2}\right)^{\frac{N}{2}} N^{\frac{1}{2}(k_1-k_2)} + \left(\frac{RSS_1}{RSS_3}\right)^{\frac{N}{2}} N^{\frac{1}{2}(k_1-k_3)} + \ldots} \tag{7.7}$$

It is clear how the relative residual sum of squares and number of parameters contribute to each model weight.

Remark 2
In the definition of BIC, sometimes a scaled version of (7.3) *will be used. As long as we scale by a positive constant, it does not affect selection of the minimum BIC-valued model.*

7.3 BIC Variants

For segmented regression models (and more general breakpoint models), there is some debate about whether more penalty is needed than the traditional BIC provides. We will give a brief overview of two modifications of BIC, which we label HOS and LWZ.

7.3.1 HOS

In [1], the authors consider the BIC in the form (7.5) (scaled by $\frac{1}{N}$). They propose a modification that has each breakpoint contributing three times to the parameter count. For example, for a breakpoint model with p breakpoints and a total parameter count of k (where k includes the number of breakpoints), the modification would use a modified parameter count, k_{HOS}, of $k + 2p$. The BIC thus takes the form

$$BIC_{HOS} \stackrel{\text{def}}{=} N\log(RSS) + (k+2p)\log N \tag{7.8}$$

This BIC formulation is developed in an asymptotic context. We consider also an intermediate modification of BIC by having each breakpoint contribute twice to the total parameter count. This is studied in [6], where it is motivated by (asymptotic) results in [11]. The BIC criterion in [11] is interesting in that the penalty term depends on the configuration of the breakpoints. See also [3], where a number of BIC variants are considered and discussed.

For segmented regression models with p breakpoints, the number of parameters (not including parameters modeling the residual distribution) is $2p + 2$. So the standard BIC is

$$BIC = N\log(RSS) + (2p+2)\log N \tag{7.9}$$

The intermediate version of HOS is then

$$BIC_{HOS} \stackrel{\text{def}}{=} N\log(RSS) + (3p+2)\log N \tag{7.10}$$

while the original version of HOS is

$$BIC_{HOS\text{-}2} \stackrel{\text{def}}{=} N\log(RSS) + (4p+2)\log N \tag{7.11}$$

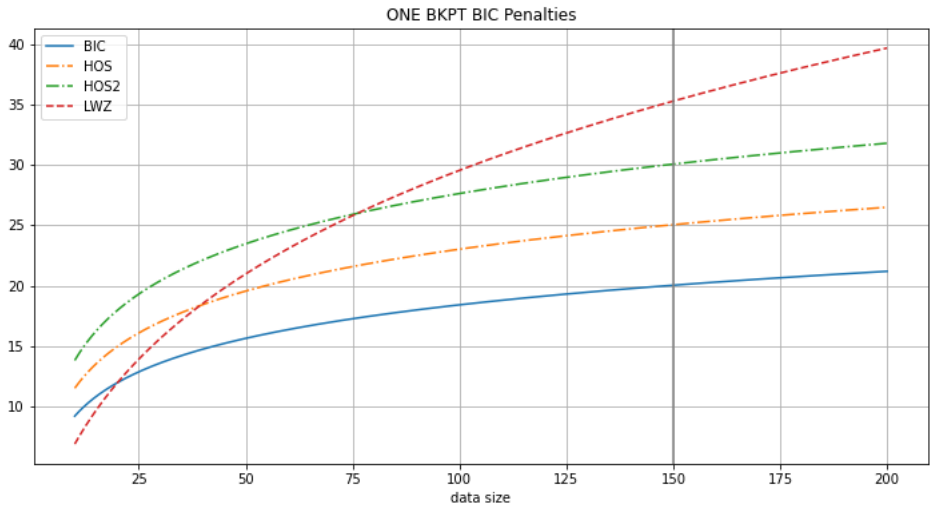
7.3.2 LWZ

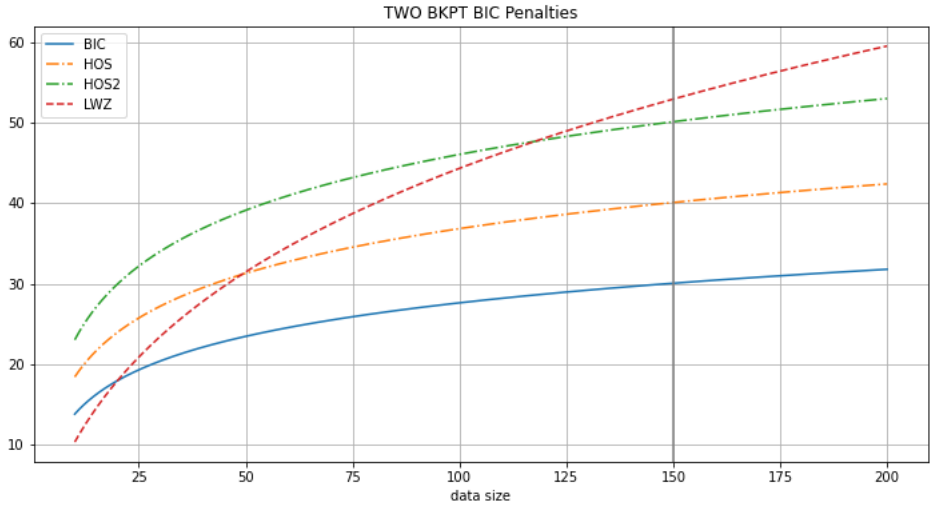
In [8], the authors propose the following modification to (7.5).

$$BIC_{LWZ} \stackrel{\text{def}}{=} N \log(RSS) + kc_0(\log N)^{2+\delta_0} \tag{7.12}$$

Based on simulation studies, they choose $c_0 = 0.299$ and $\delta_0 = 0.1$. We note that their formulation includes a degree-of-freedom adjustment which is immaterial for the segmented regression models we consider, so we do not include it.

Here we plot the penalty terms of these BIC variants. Specifically, we plot $BIC - N \log(RSS)$.





References

[1] Alastair R. Hall, Denise R. Osborn, and Nikolaos Sakkas. "Inference on Structural Breaks using Information Criteria". *The Manchester School* 81 (S3) 2013, pp. 54–81. DOI: 10.1111/manc.12017.

[2] Derek J. Hudson. "Fitting Segmented Curves Whose Join Points Have to be Estimated". *Journal of the American Statistical Association* 61 (316) 1966, pp. 1097–1129.

[3] *Joinpoint Help Manual 4.8.0.1*. National Cancer Institute. National Institutes of Health, Bethesda, Maryland, 2020.

[4] Robert E. Kass and Adrian E. Raftery. "Bayes Factors". *Journal of the American Statistical Association* 90 (430) 1995, pp. 773–795.

[5] Hyune-Jy Kim et al. "Permutation Tests for Joinpoint Regression With Applications To Cancer Rates". *Statistics In Medicine* 19 2000, pp. 335–351.

[6] Jeankyung Kim and Hyune-Ju Kim. "Consistent Model Selection in Segmented Line Regression". *J Stat Plan Inference* 170 2016, pp. 106–116. DOI: 10.1016/j.jspi.2015.09.008.

[7] Helmut Küchenhoff. "An exact algorithm for estimating breakpoints in segmented generalized linear models". *Computational Statistics* 12 1997, pp. 235–247.

[8] Jian Liu, Shiying Wu, and James V. Zidek. "On Segmented Multivariate Regression". *Statistica Sinica* 7 1997, pp. 497–525.

[9] Adrian E. Raftery. "Bayesian Model Selection in Social Research". *Sociological Methodology* 25 1995, pp. 111–163.

[10] Binbing Yu et al. "Estimating joinpoints in continuous time scale for multiple change-point models". *Computational Statistics and Data Analysis* 51 2007, pp. 2420–2427.

[11] Nancy R. Zhang and David O. Siegmund. "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data". *Biometrics* 63 2007, pp. 22–32. DOI: 10.1111/j.1541-0420.2006.00662.x.