TESTING FOR STRUCTURAL BREAKS IN THE EVALUATION OF PROGRAMS

Suzanne J. Cooper
John F. Kennedy School of Government, Harvard University

Anne Morrison Piehl
John F. Kennedy School of Government, Harvard University and NBER

Anthony A. Braga
John F. Kennedy School of Government, Harvard University

David M. Kennedy
John F. Kennedy School of Government, Harvard University

April 2001

Keywords: *parameter stability, program evaluation, youth homicide*
JEL No. C22, K42, I12

The Boston Gun Project was supported under award #904-IJ-CX-0056 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Piehl appreciates the support of the Robert Wood Johnson Foundation. Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice. The authors appreciate the helpful comments of two anonymous referees and the editor, as well as Christopher Avery, Richard Blundell, Kristin Butcher, David Card, James Stock, and seminar participants at the University of California, Berkeley, Harvard University, Yale University, and the NBER. Please address correspondence to Anne Piehl, 79 JFK Street, Cambridge, MA 02138, *anne_piehl@harvard.edu*.

TESTING FOR STRUCTURAL BREAKS IN THE EVALUATION OF PROGRAMS

Abstract

A youth homicide reduction initiative in Boston in the mid-1990s poses particular difficulties for program evaluation because it did not have a control group and the exact implementation date is unknown. A standard methodology in program evaluation is to use time series variation to compare pre- and post-program outcomes. Such an approach is not valid, however, when the timing of a potential break is unknown. To evaluate the Boston initiative, we adapt from the macroeconomics literature a test of unknown break point to test for a change in regime. Tests for parameter instability provide a flexible framework for testing a range of hypotheses commonly posed in program evaluation. These tests both pinpoint the timing of maximal break and provide a valid test of statistical significance. We evaluate the results of the estimation using the asymptotic results in the literature and with our own Monte Carlo analyses. We conclude there was a statistically significant discontinuity in youth homicide incidents (on the order of 60 percent) shortly after the intervention was unveiled.

I. Introduction

An initiative to reduce youth homicide citywide was unveiled in Boston in the summer of 1996. The "pre/post" comparison is impressive: the number of young victims of homicide fell from 3.8 per 100,000 in 1995 to 1.25 in 1997. Yet a scientifically valid evaluation requires more than the test of difference in a simple time series. The usual approach in such settings is to compare the difference-in-differences. But in this example, because of the design of the program, the precise implementation date is unknown. Perhaps more important, there is no appropriate control or even comparison group, as all those people and locations at high risk were targeted by the program. As a result, neither difference can be computed.

Given that a difference-in-differences research design is not available, we turn to the macroeconomic literature on unknown break points for guidance. This literature has concerned itself with identifying regime changes using time series data, which seems appropriate for evaluating an initiative which aimed to change the regime under which youth functioned on the streets of Boston. The key advantage to the unknown break point approach is that the break might not be where one thinks it is. In particular, implementation lags in policy interventions make it very difficult to say that a particular date defines the break between the "pre-program" and "post-program" periods, even when the implementation date is known. In such settings, if a researcher were to use the time series to judge where to time the break, even informally, the test would not have the correct critical values. Rather, statistical significance in a traditional pre/post analysis is overstated if ex post information is used to locate the time when the program is said to have begun.

In this paper we build on the literature on unknown break points, applying it in a new setting, program evaluation. We use the methods to evaluate the Boston Gun Project and the

intervention it generated, Operation Ceasefire. This setup allows us to report new results which (1) test in the most general way for the existence of a break in the data series and (2) pinpoint the timing of the break. From these results we can calculate the magnitude of the impact. In addition, we explore alternative counterfactuals in order to assess the validity of the interpretation of the break as a program effect. Furthermore, the existing time series literature provides only asymptotic distributions for the statistics used. We conduct Monte Carlo analyses to give us some insight into the use of this technique with finite samples and count data.

The use of time series variation and the unknown break point technique has broad applicability to other instances of program evaluation. In particular, appropriate comparison groups often cannot be located when the program target is at an aggregate level. This paper discusses ways in which the technique can be applied to other program evaluation cases with possibly different characteristics.

II. The Boston Project

Before turning to the development of the methodology, we describe the essential elements of the Boston youth violence reduction initiative. In the early 1990s, amid concern about the growing numbers of youth involved in homicides, both as victims and as offenders, a multi-agency working group was convened to research Boston's youth violence problem, to craft a strategy to respond to the conditions, to implement that strategy, and to evaluate the experience.[1] Throughout the effort, the goal of the Boston Gun Project (BGP) was to reduce

[1] Membership in the working group included representatives from the Boston Police Department, local and federal prosecutors, the Bureau of Alcohol, Tobacco and Firearms,

youth homicide in the city of Boston in the relatively near term. Given that the agencies represented in the working group had responsibility for the city as a whole, it was not possible to designate a priori a portion of the city as a control site. Rather, the group was prepared to work against youth violence wherever it appeared within the city limits.

Research undertaken by the group indicated that youth homicides were concentrated geographically (most incidents occurred within the three neighborhoods of Roxbury, Dorchester, and Mattapan) and demographically (most victims were male and African American). Furthermore, an incident-by-incident analysis revealed that, conservatively, a minimum of 60% of the homicides could be characterized as arising out of a nexus of disputes across "gangs." (Relative to gangs in other cities, Boston gangs were relatively small with somewhat loose organization.[2]) These gangs were involved in longstanding antagonisms with each other, which sometimes erupted in series of retaliations. These disputes were not necessarily confined geographically. Because of the dynamics of these disputes, it was not possible to designate "control" sites for the intervention.[3]

---

probation and parole departments, community outreach workers, and academics.

[2] For example, the high degree of corporatization reported by Levitt and Venkatesh [2000] was not the norm observed in Boston.

[3] Here we discuss the conceptual impossibility of utilizing control sites. It would also have been difficult or impossible in practice. It is highly unlikely that any of the agencies involved in the project would have agreed to "set aside" certain sections of the city as control sites for the purpose of improving the evaluation design.

The majority of the youth homicide incidents were even further concentrated among a group of young men with whom many in law enforcement were familiar due to criminal activity and gang associations, which are generally advertised through clothing or other symbols.[4] The working group designed a strategy to target deterrence messages and sanctions toward those individuals and groups most active in perpetrating violence, and to do so on an ex ante basis. This was a departure from the usual practice of concentrating enforcement efforts on investigation after crimes occurred. From the point of view of prevention, there were two key premises for the strategy. First, "key personnel" were identifiable. Second, retaliations were responsible for an appreciable amount of the violence. If key gangs and key individuals within those gangs could be identified and reached by a deterrence message, the inter-gang dynamic could be interrupted. As is true in the general tipping model, if the interruption of the dynamic were substantial enough, there could be a dramatic shift in the level of violence. In short, the regime on the streets could be moved from a "high violence" equilibrium to a "low violence" equilibrium.[5]

Based on these premises, an intervention known as Operation Ceasefire was developed. Because law enforcement resources in the city were limited, choices had to be made as to which incidents and threats could be addressed. One element of the intervention involved members of the working group sharing information to identify those involved in violent disputes in order to

---

[4] For more details on the basic research on youth homicide in Boston, see Kennedy, Piehl, and Braga [1996].

[5] This characterization was laid out ex ante [Kennedy et al. 1996], not as an explanation for the results which appear later in the current paper.

target resources. Enforcement agencies customized sanctions to individuals and gangs depending on their (often extensive) prior involvement with the criminal justice system by enforcing conditions of probation and/or parole. Due to the longstanding nature of antagonisms between gangs, practitioners could predict conflicts with some success. In such cases, both enforcement and social services mobilized to prevent retaliations. Finally, in order to deter initiations of violence and retaliations, the working group "advertised" the goals, capacities, and achievements of the initiative to individuals and gangs identified as "at risk" of violent assault or victimization.

The project evolved as follows. The working group was constituted early in 1995 and met regularly during that calendar year to research the basic problem of youth homicide. Results were presented to the Boston Police Department and other partners in January of 1996. Based on this research, the strategy to be implemented was defined and refined over the spring. In May, the first formal event was held, a forum explaining the strategy to a group of gang members. At the end of June the strategy was publicly announced to the public at a conference organized by the Boston Bar Association.[6]

Two attributes of the intervention pose particular challenges for evaluation design. First, the project evolved over time, moving from the research phase to strategy development and then to implementation. Even once the strategy was "in place," the intervention's attributes changed with the circumstances of particular violent outbursts and as the nature of youth violence in the city evolved. As a result, it is impossible to assert with confidence the date on which

[6] For a detailed time line and a more discursive description of the details of the intervention, see Piehl, Kennedy and Braga [2000].

implementation of the intervention began and even more difficult to determine the date of any effect without looking at the data. Therefore, an evaluation should treat the timing of any program effect as unknown. Second, the hypothesis of the intervention design was that affecting a dispute between two gangs would naturally have spillovers to other groups with which the original gangs feuded. The spillovers were expected to result from particular enforcement actions. Further, the working group hoped the deterrence message would be heard by others uninvolved in the original dispute.

As a result of these features of the Boston initiative, "treatment" could not be randomized over individuals. There are two reasonable alternative evaluation designs in this case: time series and panel. In a panel design, the experiences of youth in other cities provide the counterfactual for what would have been expected to occur in Boston in the absence of the program. However, trends vary greatly across cities and, in addition, there may well have been spillovers as other cities adopted aspects of the program after it received positive media attention early on. Therefore it is not obvious that using other cities is preferable to using control variables within Boston. In this paper we utilize the time series variation within Boston alone, applying a technique to locate a discontinuity in the youth homicide rate while controlling for characteristics of the city that arguably changed over time in a manner consistent with the observed changes in youth violence.

III. Descriptive Statistics

The dependent variable for the evaluation of the initiative is the monthly number of homicide victims aged 24 or under, provided by the Boston Police Department. Figure 1 plots

the raw data, from January 1992 through May 1998. The number of homicides is relatively small, with a number of months recording zero events. The series exhibits a great deal of variation, some of which is seasonal (with August and September having the highest homicide rates). This seasonal variation notwithstanding, one can see that the number of homicides is particularly low late in the time series.

Table I reports descriptive statistics by year for youth homicide, population, and several additional controls. The first column shows that the average number of youth homicides per month was between three and four in the early 1990s. The number of incidents falls to around one per month by the end of the period.

In explaining changes in youth homicide, it is potentially important to control for the size of the population at risk. The second column reports the number of African American males ages 15-24, as most victims are members of this demographic group.[7] The number in this group fell by 7 percent over the time series. It is worth noting that the results are not sensitive to the particular population control used. In fact, the population variable is generally not statistically significant in the regressions reported below.

The youth homicide rate is reported in the third column as the number per 100,000 population, as is standard in reporting crime data. This rate will not be used in the subsequent analyses, but is reported here for illustrative purposes. From 1992 to 1997, the youth victimization rate fell 56% from approximately 24 to 10 per 100,000.

---

[7] Of the 169 homicide victims aged 24 and under from 1992-1994, 72% were black males. (Authors' calculations from the Supplementary Homicide Reports.)

Table I includes values for several additional variables which will be used later as controls. The number of homicide victims aged 25 and older ("adult" homicides) also fell over this period. When divided by the population aged 25-44, the adult victimization rate fell 28% from 1992-1997. To control for the booming economy, we use the teen unemployment rate, which is reported in the final column. Unfortunately, due to small sample sizes in the Current Population Survey, this variable is only available annually, and only for the state as a whole. The teen unemployment rate fell by over half during this period.

III. Tests of Structural Change

The common practice in the program evaluation literature of defining a dummy variable for the "post-program" period and testing for a change in mean or other parameters is generally flawed on two grounds. First, the precise timing of an effect of a program is not known, even if the start date of the program itself is known. Therefore, the "post" dummy variable may not actually enter at the appropriate time for evaluating an effect of the program, and for this reason the estimated program effect may not be correct. Second, as Banerjee, Lumsdaine, and Stock [1992] and others point out, when the break point in the data is not known a priori, conventional hypothesis testing is not valid. Specifically, conventional critical values for tests of parameter change are not valid when the break point is inferred from examination of the data. If one were to use the conventional critical values, one might erroneously conclude that a break exists. In addition, a single break in a time series could lead to statistically significant Chow statistics for many surrounding data points. Thus a researcher could arrive at an erroneous conclusion regarding both timing and statistical significance of a break in the time series.

In the program evaluation setting, these issues are particularly salient. Even when the start date of a program is known, implementation lags and even leads (through "announcement effects" where the program has an effect before it even begins) make it virtually impossible to identify the timing of effect a priori. Rather, one determines when the effect occurred by looking at the data. It is this process that renders the conventional critical values for a Wald statistic invalid.

What we propose (and apply in Section IV) is to use a test for structural change in the case of an unknown break point, developed in the econometrics literature on time-series analysis. The time-series literature has addressed the issue of an unknown break point in an extensive and lengthy literature which encompasses many different approaches to this problem. However, the branch of the literature most applicable to the approach proposed for the program evaluation setting has its origins in the Quandt Likelihood Ratio (QLR) statistic [Quandt 1960]. This statistic is the maximum F-statistic from a Chow test, evaluated over all possible break points. In other words, one searches for the maximal test statistic for the test of the null hypothesis of no change in parameters. More recently, others have extended this literature[8] to allow tests of partial structural change (i.e., changes in a subset of the parameters) and Andrews [1993] has tabulated asymptotic critical values for the maximum Wald statistic.

[8] See Stock [1994] for a complete review of this literature.

The approach is as follows. Given a stationary time series,[9] define a Wald statistic for the null hypothesis that the parameters of interest do not change between periods:

$$H_0: \beta_t = \beta_0 \text{ for all t}$$

$$H_A(\pi): \beta_t = \begin{cases} \beta_1, t = 1,\ldots,T\pi \\ \beta_2, t = T\pi + 1,\ldots,T \end{cases}$$

where $\pi \in (0,1)$ is the fraction of the sample before the point of parameter change, i.e., $T\pi$ is the time of the change. There can in addition be another parameter vector δ_0 which is invariant with respect to π. In other words, one can test the null hypothesis that the parameters do not change against the alternative hypothesis that a particular subset of the parameters does change. The sup Wald statistic is therefore analogous to the QLR statistic (and therefore also to a Chow test).

[9] We applied a Dickey-Fuller test to test for a unit root in the data. We rejected the existence of a unit root and therefore can apply this procedure. While we applied the Dickey-Fuller test to the full data sample, we recognize the potential for over acceptance of the null when there is a break in the data. However, we rejected the null of a unit root, so this over acceptance is not relevant.

The Wald statistic for the null hypothesis of no change in parameters is computed for all possible break points in the time series. The maximum of these (i.e., the sup Wald) is the test statistic of interest. Clearly if one simply looked for the maximal Wald statistic over all possible breaks, conventional critical values for this test statistic could not apply since one searched for the highest possible value. Rather, alternative critical values specifically for the sup Wald, and not the Wald statistic in general, must be applied.

If the maximal Wald statistic exceeds the appropriate critical value, then one rejects the null hypothesis of no break in favor of the alternative that the parameters change at the point identified by the maximal Wald statistic.[10, 11] If this statistic does not exceed the critical value, then one can conclude that there was no statistically significant change in parameters. In the program evaluation context, a finding of a sup Wald statistic that does not exceed the critical value is interpreted as finding no program effect. It is important to note that when one does find a break it can only be indirectly attributed to the program. Although a break in the parameters of a regression relationship has been found, but it has not been proven that this break was caused by the program intervention. Of course, this limitation in interpretation is equally present in simple "pre/post" program analyses.

---

[10] In applying this unknown break point test we refer to the critical values in Andrews [1993] as well as conduct our own Monte Carlo analysis to account for our small sample size and count data.

[11] Bai [1997] derives a confidence interval for the location of the break point. However this is not applied here because in the program evaluation setting the primary goal is to determine the existence of any program effect.

There are, of course, many alternatives to the sup Wald test applied here. The CUSUM statistic [Brown, Durbin, and Evans 1975], the Nyblom [1989] statistic, and the Andrews and Ploberger [1994] approach all address the unknown break point (change point) issue in analogous but somewhat different ways. Generally these approaches evaluate the adequacy of the model in a diagnostic sense and are not specifically geared toward identifying the location of the break, although the break can be identified.[12]

One characteristic of the sup Wald test in particular is that the alternative hypothesis is well-specified. In the program evaluation setting, we have a particular alternative hypothesis in mind, i.e., that there was a change in parameters at the time of break. We prefer the sup Wald test here for several reasons: (a) our goal is larger than an assessment of model accuracy, (b) the alternative hypothesis is well specified, and (c) the sup Wald test has a particularly compelling intuition for the program evaluation case. In addition, the sup Wald test has asymptotic power against the more vague alternative hypotheses. It should be noted, however, that there is no consensus in the time series literature on which approach is most suitable in general.

In our application we expect any program impact to appear as a discrete change in mean, for reasons motivated in Section II. However, one might, in some circumstances, want to consider a gradual transition from one level of the outcome variable to another. In other

---

[12] In addition, one could consider approaches even farther afield, such as the approach of Hamilton [1989] and Potter [1995] which parameterize a regime-switching mechanism and then estimate the parameters of this mechanism. Another approach is regression tree analysis, as applied in Cooper [1998]. Even farther afield, one might consider a Bayesian approach where a prior on the break is specified and then a posterior distribution of the break is derived.

circumstances one might consider multiple discrete changes. The sup Wald test described and applied here is applicable in both of these cases. These types of models are considered both in the time series literature and in our empirical application, in Section VII below.

In practice, to apply the unknown break point technique in a program evaluation setting, one first needs to define the regression relationship of interest, i.e., the outcome that one hypothesizes might be affected by the program as well as any control variables. Second, one specifies what parameters are permitted to change, i.e., mean or trend, some subset of the regression parameters, or all regression parameters.

One decision to be made when applying this technique for searching for a data break is the choice of the "trimming" value. When one searches over all possible locations for a break in some parameters, one needs to specify how far into the sample one starts looking for a break and how close to the end of the sample one stops looking. The reason for not looking from the first observation to the last is that there must be a sufficient number of observations on either side of the point under consideration to estimate the regression relationship both before and after the break point.

Trimming is generally done symmetrically from both ends of the sample. For program evaluation, because one wants to know the answer in real time, the range in which one wants to consider a break may come close to the end of the sample. In small samples the inclination may be to trim a higher percentage of the sample in order to have sufficient observations to evaluate the earliest and latest potential break points. However, at the same time, one sacrifices having a large range of data points to consider. When evaluating a program, one needs to consider how much data to collect before attempting an evaluation. The need to trim in order to estimate

parameters of interest alerts the evaluator to the importance of accumulating enough experience (i.e., data) before testing for a break (program effect).[13]

In addition, when applying the structural break methodology to program evaluation, one must consider whether to look in a narrow "window" of dates for the break. The time series literature does not address this question. In program evaluation, however, this issue is particularly relevant because a break may be found at a date sufficiently far from the program date as to question its interpretation as a program effect.

There are two possible ways of addressing this situation. First, one can a priori construct a window or range of dates in which to look for a program effect, and restrict application of the structural break methodology to those dates. This does not raise the Chow test problem as long as the range of dates is selected by knowledge of the program and not by looking at the data. A second possibility is to allow for multiple breaks in the time series. Then when a break is found far from dates relevant to the program, one can condition on that break and search for another break to see if there is a break that could be attributable to the program. These approaches are not particularly different in the sense that in either case one must use knowledge of the program to determine where to consider a break in the time series as attributable to the

---

[13] While trimming may not be of critical importance in much macroeconomic data, it does raise an interesting point in the program evaluation setting. How much data should one collect on either side of a data point that one is not supposed to assume a priori? This question is particularly relevant because it is often costly to wait for experience to accumulate before performing an evaluation.

program. However, in both cases one can do better than by specifying a date a priori for a program effect and applying a Chow test.

The unknown break point tests in general, and the sup Wald test in particular, are broadly applicable to the program evaluation problem even though they are derived from the time series econometrics literature. One can search for and test for the existence of a break in some outcome, thus avoiding the pitfall of assuming when a program had its effect. One can define the type of effect one is looking for (for example, a change in mean or some other parameters). And most important, one achieves a statistically valid conclusion regarding the existence (or not) of a break in the time series.

IV. Empirical Results

This section tests for a structural break in the Boston youth homicide time series using the sup Wald methodology. Here, we test for a change in the mean number of youth homicides.[14] The initiative hoped to move the level of homicide to a new, lower equilibrium. That is, controlling for other factors, if there was a program effect, it should appear as a discrete shift in mean.[15] As a result, our application is relatively straightforward: regress the dependent variable on a series of controls, including month indicators and population, and test for a change

[14] As the discussion above indicated, the sup Wald test is applicable for a broad range of different specifications.

[15] There is no reason, from what is known about youth homicide, to believe that there would be changes in the seasonality or in the relationship between economic conditions, for example, and homicide.

in the constant. As a robustness check, we consider alternative specifications of the form of the break.

Because of the count nature of the data, Poisson regression is a logical choice for specifying the model. However, in the Poisson the variance equals the mean. As a result, testing for a break in mean must also be a test for a break in the variance. We did not want to test for such a compound hypothesis. Instead, we simply ran OLS, correcting the standard errors for heteroskedasticity. This allowed us to test for a change in the mean without making any implicit assumptions about the variance.

As discussed in the previous section, we must define a range of dates for possible treatment effect. Any break found outside that window is not plausibly attributable to the program. In order to be as agnostic as possible with regard to potential announcement effects or implementation lags, we chose a wide window. Therefore, we defined the window for possible impact to run from January 1996 through May 1997, giving generous room for announcement effects and 12-months post actual implementation to allow for lagged impact.

Table II reports the results of running OLS on the monthly number of youth homicides from January 1992 through May 1998 for three sets of control variables, trimming 15 percent off each end of the time series when searching for the maximal break point.[16] All models include

[16] Fifteen percent trimming is the amount recommended by Andrews [1993]. As less of the sample is trimmed off the ends, the critical values get larger in recognition of the fact that fewer data points are pinning down the ends of the time series. On the other hand, in a finite sample, trimming a greater proportion of the data can dramatically reduce the (sometimes already small) sample size. We found no reason to deviate from the recommended level of 15%

controls for the population of black males aged 15-24 and a full set of month dummies. In order to avoid introducing any breaks into the data, we linearly interpolated the data which are only available annually (i.e., population and the teen unemployment rate).[17]

The first column reports the maximum value of the Wald statistic for each model. For model A, the sup Wald value (33.70) occurred in June 1996. This test statistic far exceeds the conventional critical values for a Wald statistic, e.g., the 5% Chow critical value is 3.97 for 77 observations. Furthermore, the test statistic exceeds the 5% asymptotic critical value for a break in one parameter with 15% trimming from Andrews [1993], which is 8.85. (We would expect the Andrews critical value to be much higher than the Chow value given the search for the maximal value of the Wald statistic.) However, because of our small sample and the count data, we do not feel comfortable relying on the asymptotic critical values and therefore, in the next section, assess statistical significance with Monte Carlo results. The effect size is reported in the final column: a reduction of 2.49 homicide victims per month, or an approximately 72% decline. Note that June 1996 falls inside the window for possible treatment effect. However, before we can call this decline purely a treatment effect, we must also consider other factors that might have led to a decline in youth homicide in the absence of the program.

In model B, we used the rate of "adult" homicide victimization to control for the counterfactual for youth homicide in the absence of the program. This is quite a strict test given

---

trimming. Nonetheless, as a robustness check, we ran our specifications with 10% trimming. The month of maximal break was unaffected by changing this parameter.

[17] The results are not sensitive to the choice of month for placement of the annual value for interpolation.

that the intervention could very well have affected the victimization of older people (directly, if younger people reduced their victimization of older people,[18] or indirectly, because enforcement was targeted at a type of offending, not strictly on age). With adult homicide as a control, the break continues to be located in summer of 1996, now August, though the value of the sup Wald is now 20.93 and the effect size is somewhat lower (a 2.12 victim reduction per month). With the break so located, the adult homicide rate has a p-value of 0.12.

Model C adds the teen unemployment rate to the controls for the previous model. Adding this covariate reduced the maximal value of the Wald statistic by over half (to 8.89). Since the unemployment rate in general was falling over the time period of this analysis, these results are not surprising. The break was still placed in June 1996 and the effect size remained near 60%. At the same time, the unemployment rate was not statistically significant in explaining the number of youth homicides with the break in the maximal month. The finding that unemployment is not strongly related to crime is not unusual in the literature [Piehl 1998]. The precise timing in the relationship between unemployment and youth homicide appears to be coincidental. Unemployment was falling continuously during this time due to a general macroeconomic boom. For this reason, the unemployment rate acts much like a linear trend in this analysis. We further explore the role of a linear trend in Section VI.

---

[18] Using data from Supplemental Homicide Reports, Cook and Laub [1998, Table 5] report that young homicide offenders tend to kill people who are older than them. For killers aged 13-17, 75 percent of their victims are older than the killer and over 50 percent of victims are more than five years older than the killer.

V. Monte Carlo Results

As noted earlier, the asymptotic critical values may not be appropriate for determining the statistical significance of the Wald statistics reported in Table II. Because our sample size in this paper is only 77 monthly observations, we need to generate our own critical values through Monte Carlo analysis. A second reason to do our own Monte Carlos is that the data in this paper are counts of youth homicide and critical values were not calculated for count data in the published literature.

In order to produce appropriate small-sample critical values based on count data for comparison with our sample test statistics, we generated data using the properties of the actual data, but without a break, and then computed the size of the test for structural break. In particular, we conditioned on the independent variables in the regression (i.e., we ran a separate Monte Carlo analysis for each of the specifications reported in Table II). We conditioned on the independent variables because it is not possible to generate time series data with the properties of the actual data for the right-hand-side variables because these are measured over too short a time span to accurately determine the time series properties.

Conditioning on the independent variables is valid as long as they are truly exogenous. Exogeneity of population is difficult to refute: the number of homicides is so small relative to population that population cannot be determined by the number of youth homicides with any lead or lag. The adult homicide rate is perhaps more questionable. Exogeneity requires that the adult homicide rate not be determined by the youth homicide count (the dependent variable in our analysis). One might argue that a change in the youth homicide count could affect the adult rate in a future period. However, we consider the independent variables to be exogenous and

condition on these variables in the Monte Carlo analyses of size and power. If one is concerned about lack of exogeneity of the adult homicide rate, one would prefer specification A over specification B.

To evaluate the statistical significance of this sample test statistic using the Monte Carlo analysis, we generated data with the full-sample properties but without a break in mean. We can then see how often the test of structural break finds a break when there is none. In addition, we used 15% trimming and a test for break in mean in the Monte Carlo analysis, just as with the actual data. Monte Carlo results for each specification of independent variables were generated from 10,000 draws under these conditions.[19]

The results of the Monte Carlo analysis of size are reported in Table III, together with the asymptotic critical values for the max Wald statistic from Andrews [1993]. As is to be expected, the critical values for any particular size, regardless of specification, are higher than the conventional Chow critical values.[20] The Monte Carlo critical values are fairly consistent across

---

[19] In particular, for each specification we ran a Poisson model and computed the variance of the residuals. For size calculations we used the variance over the full period and for power calculations we used the variance in each of the two subperiods defined by the appropriate break point from the actual data for the specification. For each of 10,000 draws we used these variances to generate 77 observations using the Poisson distribution. For each draw we then ran OLS with the appropriate controls for that specification, calculating the maximum Wald statistic and its location in the time period.

[20] For 77 observations these are: 2.77 at the 10% significance level, 3.97 at 5%, and 6.98 at 1%.

specifications. There is some variation, however, which is not surprising as the number and nature of independent variables changes.

In comparing our sample test statistics from Table II with the small-sample and count-data-consistent critical values in Table III we find quite strong evidence for the statistical significance of the structural breaks found. For specification A, the sample test statistic is 33.7, which exceeds the critical value at even the 99% level (15.7). For specification B, the sample test statistic is 20.9, which also exceeds critical values beyond the 99% level (15.9). The sample test statistic for specification C is much lower, at 8.89. This lower test statistic is likely due to the decline in the unemployment rate over this time period, which absorbs some of the decline in the youth homicide count. However, as explained earlier, we do not find this a convincing explanation for the decline in youth homicide and therefore are doubtful that this decline in the sample test statistic reflects a less significant program effect. However, to be complete, we can compare this test statistic with the critical values for specification C and find that we reject the null hypothesis of no structural break with a p-value of 0.13.

For specifications A and B, the power of the structural break test is clearly not an issue since we found convincing evidence of structural breaks. However, for specification C, where there is weaker evidence of a break, we might wonder whether this is really less evidence of a break or simply low power of the maximum Wald test. Table IV reports results of Monte Carlo analysis of power of the test for the same three specifications we have considered throughout. Data were generated under the specific alternative hypothesis that there is a break in mean which matches both the magnitude and location of the break in the actual data. Clearly specifications A and B have quite high power to detect breaks. The power for the third specification is quite a bit

lower, however, reconfirming our hypothesis that it is harder to find a break, even when one exists, with this set of control variables. In particular, Table IV indicates that for specification C (i.e., controlling for population, adult homicide rate, and the questionable youth unemployment rate), we would find a break in mean youth homicide counts only 45% of the time (using a 90% critical value) even when such a break truly exists. For the 95% critical value we would find a break in mean only 33% of the time when in fact such a break exists. Therefore we should not be surprised that we reject the null hypothesis of no structural break at only the 13% level with this specification.

Overall, the Monte Carlo analysis is quite informative. Because the resulting critical values are higher than the asymptotic values, the need for Monte Carlo analysis is apparent and recommended in general when applying the test of structural break. Other interesting findings from the Monte Carlo analysis shed further light on the properties of the test. In particular, the distribution of test statistics is quite nonlinear, with the critical values rising steeply in the upper tail. We do not find this surprising since in a test of structural break we expect that spurious breaks with high test statistics will be found, but rarely, indicating that very far in the tail there will be spurious high values of the maximal Wald statistic.

Interestingly, the distribution of the location of maximal Wald statistic is not entirely uniform as one might expect when the generated data have no break. Figure 2 shows the distribution of "months" in which the maximal test statistic is found. While this figure is based on specification B, its qualities are representative. The figure indicates that the test of structural break is more likely to find the maximal Wald statistic near the ends than in the middle of the time series. While this might be of some concern in an analysis where the suspected break is

indeed near the ends of the time series, it is not of concern to us. The maximal test statistic in our sample data, as reported in Table II, is month 54 which falls in the range of months where relatively fewer maximal test statistics are found. While Figure 2 does suggest a tendency for the structural break test to find maximal test statistics near the edges of the time series, it is not the case that such findings would necessarily be statistically significant given the results in Table III, and therefore may not be of much concern overall.

VI. Robustness

The methodology presented and applied in this paper can tell us whether there is a break in the time series for some outcome variable, and if so, where that break is. In this section we consider alternative explanations for the change itself. In particular, in order to interpret a break in the time series as indicating a program effect we must be certain that the break does not appear for any other reason. As with any program evaluation question, regardless of methodology for estimating the program effect, the effect can be interpreted as due to the program only when all other possible reasons for the change have been ruled out. These alternatives will in general be specific to the problem under investigation.

One possible criticism of the conclusion of the previous section that youth homicide fell substantially is that youth homicide was falling over the entire time period in Boston and no abrupt change in the number of homicides would be found if we in fact controlled for this longer term decline. We chose specification B as a baseline because we feel that the adult homicide rate is the best control for the counterfactual for youth homicide. The first row of Table V presents the result of expanding specification B to include a linear trend control. Adding a linear

trend to this specification reduces the maximum Wald statistic to 6.93 (from 20.93). The month of maximum break is June 1996, which is substantively the same as the results in Table II.[21] The maximum Wald is much lower with the linear trend included, and in fact we reject the null hypothesis of no break at only the 26% level.

While the general impression of the program evaluation is the same with the trend included, the substantial decline in significance level requires a closer look at these models. In applying the structural break methodology, we need to have the correct specification of the regression model under the null hypothesis of no break. If a linear trend belongs in the model, then we would need to include it in the analysis in order to have the correct inference on the break in mean. However, if there is no trend in the true relationship and we include a trend variable, this will obscure inference on the break in mean. In particular, while inclusion of an irrelevant variable in ordinary regression analysis does not bias the coefficients, the structural break methodology testing for a change in mean is looking for a change in the dependent variable over time. Inclusion of a trend could absorb some of the change in mean, even if there is no trend in the data, i.e., if there are just two regimes with different (constant) means.

To understand the effect of including a trend, consider data with a break in mean but no trend. Fitting a trend to that time series will clearly absorb some of the change in mean, leaving less to identify as a break. In particular, to analyze this effect in our example, we generated data from the parameters of specification B (i.e, consistent with the data generated for the power

[21] The result is slightly different from specification B which had the maximum break in August 1996 in Table II, but specifications A and C had the maximum break in June 1996, so the general finding is upheld.

analysis of specification B presented in Table IV) with a break in mean but no trend. We then computed a power-like calculation (based on 10,000 draws as in the Monte Carlo analysis) where we considered how often we would reject the null of no break when we included a trend as a control variable even though one was not present in the data generation. Our finding is quite striking. We reject the null hypothesis of no break (even though one exists) only 48% of the time for the 90% critical value. By comparison, the power associated with specification B for this critical value is 87%. Therefore, inclusion of the trend, even when none is present in the generated data, reduces almost in half the possibility that a known break can be found. For those cases when the break month is within 6 months of the actual break date, the ratio of the maximum Wald statistic for the case with trend relative to the case without trend is only 0.62.[22]

Clearly inclusion of a trend, even when none is present in the data, substantially reduces the possibility of finding a break, even when one does exist. The trend results confirm our interpretation of specification C, which showed weaker evidence of a break in mean when the (downward trending) unemployment rate was included. The fact that even when controlling for a linear trend the timing of break is virtually the same, although with lower statistical significance, reinforces our finding that there was a dramatic decline in youth homicide in Boston in the middle of 1996 that was not part of an otherwise existing downward trend in youth homicide. The broader point is that specification of trend (and trending variables) must be done carefully when using the structural break methodology.

---

[22] Of the break months found in the Monte Carlo analysis, 82% were within 6 months of the generated break date when no trend was included, relative to only 66% when a trend was included.

To further investigate the impact of including the trend control in the analysis, we considered the possibility of a break in that trend, as well as a break in mean. The result of this specification is presented in the second row of Table V. In this case we find that the timing of break is roughly similar (in fact it is identical to specification B and within 2 months of all other specifications). The statistical significance of the break (now a break in two parameters) is also somewhat improved from the case of including a trend with only a break in mean. The null hypothesis of no break in mean and trend is rejected at the 12% level using the Monte Carlo critical values. This specification in fact resembles specification B more than it resembles the previous specification (which was B plus a trend). Our interpretation is that including the linear trend absorbed some of the break in mean. Allowing the trend to break reduced the absorption of the mean decline by relaxing the restriction that the trend be linear. The fact that the trend broke suggests that a linear trend was mis-specified.

Overall we conclude that the change in the mean monthly number of youth homicides in Boston was not due to an overall downward trend in the data.[23] The results of including the linear trend, as well as allowing a break in both mean and trend, give qualitatively similar findings to specification B. Furthermore, neither of these alternative specifications leads us to question specification B as an appropriate model. By controlling for a general downward trend, as well as the decline in adult homicide that occurred at the same time and probably even due to

[23] We find similar results when we use the national youth homicide rate to represent the counterfactual instead of the linear trend. An analysis of the youth homicide experience of other large U.S. cities over a similar time period revealed no robust or systematic breaks. Together, these results suggest the Boston experience was unique.

the same program, we are confident that we have ruled out any likely possible alternative explanations for the decline in the youth homicide rate.

VII. Extensions

A final consideration in applying the structural break methodology to program evaluation questions is the specification of the form of the possible break. In general it is possible to have a program that suggests a more complicated program effect than a break in mean. If one thought there were going to be a gradual implementation, a break in mean and trend could be allowed. Alternatively, one might think of implementation in stages, allowing multiple points of impact, in which case multiple changes in mean could be allowed. These generalizations have been developed in the time-series literature.[24]

We do not expect these more complex specifications of the program effect to help explain the Boston Gun Project because the program was anticipated to lead to a change in mean if it worked at all. Rather, we apply these generalizations of the form of program effect to illustrate the flexibility of using the break-point approach for program evaluation. The first possibility has already been considered. The break in mean and trend specification discussed in the previous section allows for the possibility of a shift in mean at the time of program effect, with a different trend from that point forward. If we thought that the program would have an immediate downward effect on youth homicide, as well as a gradual decline in the subsequent period, this specification would make sense. As discussed earlier, these results are presented in

[24] For example, Andrews [1993] allows changes in multiple parameters and Bai and Perron [1998] analyze multiple breaks.

the second row of Table V. The results of this specification do not suggest a gradual decline (or a gradual return to the previous level) of youth homicide. The results on the break in trend suggest that this variable is simply acting as an alternative means of describing the change in mean that we observe.

Another alternative specification of a program effect could be an effect in stages. By looking for only one break in mean we have not allowed for the fact that a subsequent change in mean might occur. We implement this alternative in our context and the result is presented in the third row of Table V. The Bai and Perron [1998] approach involves searching for an initial break and then conditioning on that break and then searching and testing for a second break. When we allow for a second break in specification B, conditioning on the first break, the maximum Wald for the second break is only 5.42 in November 1996. This value is substantially below the asymptotic critical value at the 10% level and therefore the null hypothesis of no second break, conditional on the first, is not rejected.[25] Considering the multiple break and break in mean and trend results together, it does not appear that the decline in youth homicide in Boston took place in discrete stages or as a gradual decline.

Finally, although also not relevant in our application, one could also allow the relationship between the independent and dependent variables to change as a result of a program. For example, one might think that the underlying relationship between unemployment and job-seeking behavior could change with a particular type of training program. One can use the structural break methodology, testing for a change in a particular coefficient in addition to (or instead of) a change in mean and/or trend.

---

[25] We do not do a Monte Carlo analysis of multiple breaks due to the small sample size.

Overall, our analyses of alternative specifications of the form of the break do not suggest any different findings on the existence of a program effect in our application. There do not appear to be multiple stages of the decline in youth homicides, nor does there appear to be any substantive difference in the findings even if we allow a break in trend to permit a gradual transition. Indeed, a gradual transition following the commencement of the intervention does not appear to have taken place.

A break in mean following implementation is the clearest description of what the data show. In addition, we interpret this decline in mean as supporting our view that the decline is due to the BGP program. If any alternative forms of break gave different results we might wonder how to interpret the findings. But all specifications point to a decline in mean within three months of the best approximation of implementation. Together with the robustness analysis in Section VI, we have confidence that we have estimated the effect of the BGP program to be a decline of 60% in youth homicide, which translates into two fewer youth deaths per month.

VIII. Conclusion

Substantively, this evaluation has found that there was a statistically significant break in mean associated with substantial decreases youth homicide in the summer of 1996. This discontinuity coincides with when the BGP was implemented. Controlling for population, the adult homicide rate, and a linear trend, we have confidence that we have captured a program effect rather than an unrelated change in youth homicide. Any alternative explanation for the drop in youth homicide over this period would need to be able to account for the suddenness of

the change at that time, and we have not discovered any convincing explanations with this quality. As with any program evaluation, there is necessarily some uncertainty about which particular attributes of the program were necessary to the impact. We are much more confident in our conclusion of the existence of a program effect having used statistical methods that take into consideration the unknown timing of the break point than if we had used traditional program evaluation methods.

From applying a test of structural break to program evaluation, we conclude that the method is flexible and easy to use. While our application tested for a break in mean because the model underlying the intervention was a tipping one, many other types of effects can be evaluated with this procedure. One can test for a break in the entire set of controls ("regime shift") or, as more likely relevant to program evaluation, for a break in trend or a break in the relationship of the outcome to a single control variable. The structural break test is useful because it can identify timing and statistical significance of program effects even when the timing of effect is uncertain a priori, and can give different inference from the usual methods. The last point is not a technical detail. Traditional Chow tests overstate statistical significance when used for program evaluation. Given that the primary motivation for evaluating a program is to test whether an intervention "worked," using appropriate methods for statistical inference is essential.

References

Andrews, Donald W. K. (1993), "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, vol. 61, no. 4 (July), pp.821-856.

Andrews, Donald W. K. and Werner Ploberger (1994), "Optimal Tests When a Nuisance Parameter is Present Only Under the Alternative." *Econometrica* 62.6 : 1383-1414.

Bai, Jushan (1997), "Estimating Multiple Breaks One at a Time," *Econometric Theory*, vol. 13, pp.315-352.

Bai, Jushan and Pierre Perron (1998), "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica* 66.1 : 47-78.

Banerjee, A., R .L. Lumsdaine, and J. H. Stock (1992), "Recursive and Sequential Tests of the Unit Root and Trend Break Hypotheses: Theory and International Evidence," *Journal of Business & Economic Statistics*, vol. 10, pp.271-287.

Brown, R.L., J. Durbin, and J.M. Evans (1975), "Techniques for Testing the Constancy of Regression Relationships over Time with Comments," Journal of the Royal Statistical Society, Series B, 37, pp.149-192.

Cook, Philip J. and John H. Laub (1998), "The Epidemic in Youth Violence," in Michael Tonry and Mark H. Moore, eds., *Youth Violence*, pp.27-64.

Cooper, Suzanne J. (1998), "Multiple Regimes in U.S. Output Fluctuations," *Journal of Business & Economic Statistics*, vol. 16, no. 1 (January), pp.92-100.

Hamilton, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, vol. 57, pp.357-384.

Kennedy, David M., Anne M. Piehl, and Anthony A. Braga (1996), "Youth Violence in Boston: Gun Markets, Serious Youth Offenders, and a Use-Reduction Strategy," *Law and Contemporary Problems,* vol. 59(1), pp.147-183.

Levitt, Steven D. and Sudhir Alladi Venkatesh (2000) "An Economic Analysis of a Drug-Selling Gang's Finances," *Quarterly Journal of Economics* vol. 115, pp.755-89.

Nyblom, J. (1989), Testing for the Constancy of Parameters Over Time," Journal of the American Statistical Association, vol. 84, pp.545-549.

Piehl, Anne Morrison (1998), "Economic Conditions, Work, and Crime," in Michael Tonry, ed., *Handbook on Crime and Punishment*, Oxford University Press, pp.302-319.

Piehl, Anne Morrison, David M. Kennedy, and Anthony A. Braga (2000), "Problem Solving and Youth Violence: An Evaluation of the Boston Gun Project," *American Law and Economics Review*, vol 2(1), pp.58-106.

Potter, S. M. (1995), "A Nonlinear Approach to U.S. GNP," *Journal of Applied Econometrics*, vol. 10, pp.109-125.

Quandt, Richard E. (1960), "Tests of the Hypothesis That a Linear Regression System Obeys Two Separate Regimes." *Journal of the American Statistical Association,* vol. 55, pp.324-330.

Stock, James H. (1994), "Unit Roots, Structural Breaks and Trends." *Handbook of Econometrics*. Ed. R.F. Engle and D.L. McFadden. Vol. 4. New York: Elsevier Science, pp.2740-2841.

Table I. Descriptive Statistics: Means
(standard errors)

	Youth Homicides	Population Black Males 15-24	Youth Homicide Rate	Adult Homicides	Population 25-44	Adult Homicide Rate	Teen Unem-ployment Rate
1992	3.083 (0.543)	12,977	23.83 (4.249)	3.250 (0.676)	228,465	1.420 (0.296)	20.2
1993	4.000 (0.537)	12,455	32.00 (4.297)	4.167 (0.458)	227,218	1.833 (0.202)	18.8
1994	3.167 (0.520)	12,272	25.75 (4.220)	3.917 (0.802)	226,611	1.723 (0.352)	15.9
1995	3.833 (0.716)	12,222	31.45 (5.882)	4.167 (0.815)	231,367	1.806 (0.353)	14.7
1996	2.083 (0.358)	11,895	17.41 (2.984)	2.667 (0.512)	230,744	1.156 (0.222)	13.8
1997	1.250 (0.351)	12,038	10.37 (2.909)	2.333 (0.355)	228,696	1.020 (0.155)	12.6
1998	0.800 (0.374)	12,359	6.50 (3.031)	1.400 (0.748)	224,471	0.620 (0.331)	8.8

Sources: Homicide data were provided by the Boston Police Department. Population data came from the Bureau of the Census and unemployment rates for teens in Massachusetts were provided by the Department of Employment and Training (unpublished data).

Notes: The homicide data are monthly. The population and unemployment variables are reported annually, and have been linearly interpolated. 1998 contains data only through May. The homicide rates were calculated per 100,000 population using the associated population denominator, linearly interpolated.

Table II. Parameter Instability in Youth Homicide: Tests for Break in Mean Various Sets of Control Variables

Model	Maximum Wald Statistic	Month of Max.	Population	Adult Homicide Rate	Teen Unemp. Rate	Effect Size
A	33.70	June 1996	yes	–	–	-2.49 (72%)
B	20.93	August 1996	yes	yes	--	- 2.12 (62%)
C	8.89	June 1996	yes	yes	yes	- 2.02 (59%)

Sources: See Table I for descriptions of the variables.

Notes: 11 month indicators are included in all specifications in addition to the controls noted in columns (4) through (6). N = 77 months, January 1992 through May 1998.

Table III. Monte Carlo Critical Values: Size

	Significance Level				
	p=.20	p=.15	p=.10	p=.05	p=.01
A	7.08	7.86	8.98	11.01	15.72
B	7.10	7.93	9.01	10.90	15.86
C	7.73	8.57	9.75	11.76	17.04
Asymptotic	--	--	7.17	8.85	12.35

Notes: Monte Carlo critical values result from 10,000 draws of 77 observations from Poisson with variance from the true data, with 15% trimming, testing for break in mean only.

Asymptotic critical values are from Andrews [1993]. Critical values are for 15% trimming and a break in one parameter. Note that these critical values are not for count data.

Table IV. Monte Carlo Calculations: Power

	Significance Level			
	p=.20	p=.15	p=.10	p=.05
A	99.2%	98.9%	97.8%	95.1%
B	93.3%	90.8%	87.0%	78.9%
C	60.3%	53.8%	45.2%	32.9%

Notes: Power is for the associated critical value from Table III.

Table V. Parameter Instability in Youth Homicide: Robustness Checks

Model	Maximum Wald Statistic	Month of Max.	Asymptotic Critical Values p=.10	Asymptotic Critical Values p=.05	Monte Carlo Critical Values p=.10	Monte Carlo Critical Values p=.05
B + Trend	6.93	June 1996	7.17	8.85	9.77	11.88
Change in Mean & Trend	13.24	August 1996	10.01	11.79	14.46	17.41
Multiple Breaks in Mean	5.42	Aug96 /Nov96	9.56	11.14	--	--

Notes: N = 77 months, January 1992 through May 1998. Independent variables included in each specification are: population, adult homicide rate, trend (in the first two rows), and 11 month indicators.

For the first two models, asymptotic critical values were taken from Andrews [1993]. For the final model, asymptotic critical values were take from Bai and Perron [1998]. Monte Carlo critical values result from 10,000 draws of 77 observations from Poisson with variance from the true data, with 15% trimming.

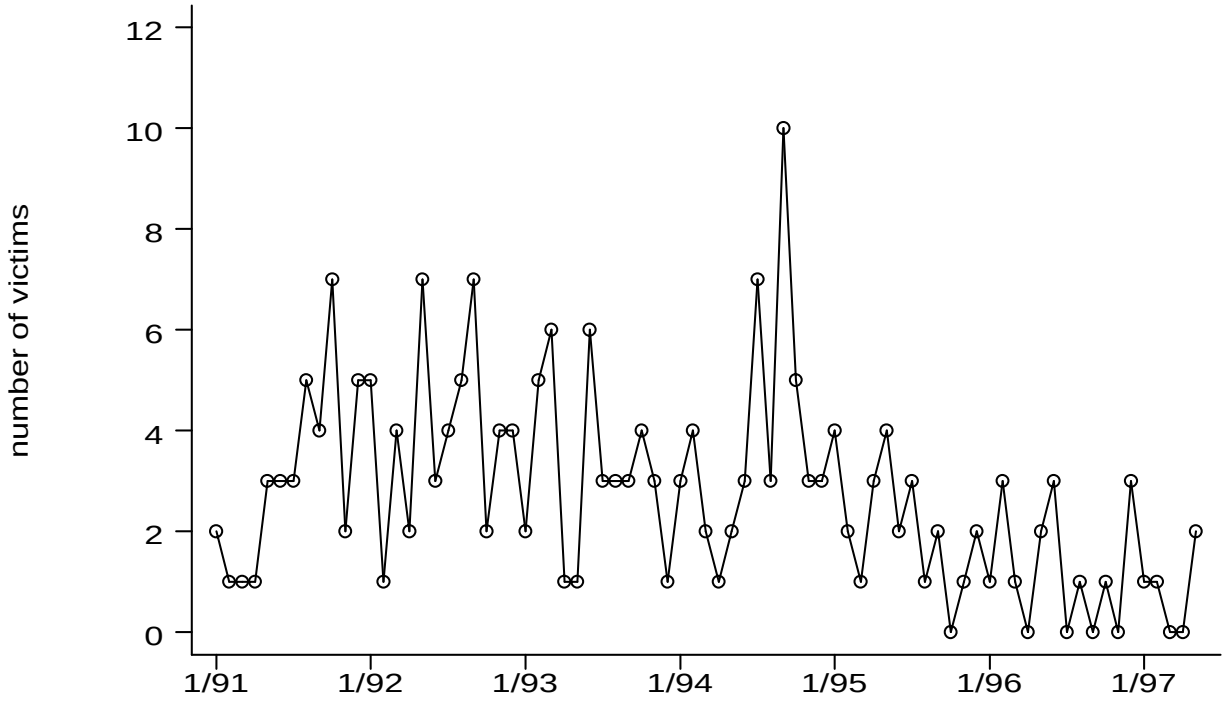

Figure 1. Monthly Youth Homicide Count 1/92-5/98

Figure 2: Monte Carlo Maximal Break Points: Size (Specification B)