CENTER FOR LABOR ECONOMICS
UNIVERSITY OF CALIFORNIA, BERKELEY
WORKING PAPER NO. 15

Testing for Structural Breaks in the Evaluation of Programs

Anne Morrison Piehl
University of California--Berkeley, Harvard University, and NBER

Suzanne J. Cooper, David M. Kennedy, and Anthony A. Braga

Harvard University

April 1999

Abstract

There are many instances of program evaluation in which the analysis of time series variation provides the best test of impact. Techniques developed for the analysis of the macroeconomy can be successfully applied to these settings. Tests for parameter instability developed in Andrews (1993) provide a flexible framework for testing a range of hypotheses. Furthermore, these tests help pinpoint the timing of maximal break and provide a valid test of statistical significance, which is particularly useful when the start date of the intervention and any effect is unclear and possibly endogenous due to implementation lags. These tests are applied in an evaluation of the effects of a comprehensive effort to reduce youth homicide in Boston in the mid 1990s. The intervention was associated with about a 60 percent decline in youth homicide.

JEL Codes: C22, K42
Keywords: *structural break, program evaluation, youth homicide*

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 Christopher Avery, Richard Blundell, Kristin Butcher, David Card, and seminar participants at the University of California, Berkeley.
Please address correspondence to Anne Piehl, 140 Warren Hall, University of California, Berkeley, CA 94729-7360, *piehl@socrates.berkeley.edu.*

I. Introduction

The Boston Gun Project was a citywide, interagency effort to reduce youth homicide in the city of Boston. Due to the dynamics of youth violence and the organization of law enforcement and social service agencies, the target unit of the effort was the city, not individuals. The media and others pronounced the effort a success on the basis of post-program results alone. How should researchers rigorously evaluate such a program? While evaluation techniques for programs aimed at individuals are well developed and the performance of these techniques in the real world well evaluated, there is a gap when it comes to evaluating programs that are intended to operate at a more aggregate level. Given the current popularity of such efforts (e.g., Healthy Cities and needle-exchange programs), there are a number of settings in which time series data provide the best test of impact. Yet the program evaluation literature has not incorporated all of the lessons time series econometrics provides.

Within time-series approaches to program evaluation, a standard methodology is to compare pre- and post-program outcomes. Use of a dummy variable for commencement of the program can be flawed. In many cases the researcher does not know the precise timing of the intervention, and may be even less confident of the timing of the effect of the intervention, if any. The time-series literature notes that when the timing of a break in a statistical relationship is data dependent (i.e., can be determined only by looking at the data), then the usual distribution of the test statistic which assumes exogenous timing of the break is no longer valid.

Techniques developed for the analysis of the macroeconomy can be successfully applied to these settings. This paper discusses the applicability of the technique of Andrews (1993) for locating and testing for a break point in a program-evaluation setting. The technique is widely applicable, allowing for tests of change in all or any subset of parameters of a regression relationship. In addition, these tests help pinpoint the timing of a break, which is particularly useful when the start date of the intervention is unclear due to implementation lags. The tests applied here, which account for the data dependence and uncertainty regarding exact timing of break point, are therefore broadly applicable in a program-evaluation setting.

We apply this technique to the task of assessing the impact of the Boston youth homicide initiative. The program was implemented over time, and therefore the precise timing of

intervention is unknown. In addition, the effect of the intervention could arguably have appeared over a wide time span. Therefore, the timing of the any effect on youth homicide can be determined only by looking at the data. For this reason, the Andrews technique is appropriate here. The dependent variable is the monthly youth homicide count in the city for the period January 1992 through May 1998. In considering whether youth homicide has fallen in response to the program intervention, we control for population changes, as well as the unemployment rate, and the rate of other violent crimes.

We conclude that there was indeed a reduction in the rate of youth homicides following introduction of the program. Specifically, we find a statistically significant decline of about 60 percent in the rate of youth homicide, after controlling for these other factors. In addition, the timing of this change in youth homicide rate is found to be within a few months of when the program was thought to have begun, and is robust to changes in control variables and time range of data. Thus, the technique is shown to be particularly applicable in the program evaluation setting, identifying the timing of effect of the program as well as allowing a valid test of statistical significance of the effect.

II. 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 outcomes either in the mean or through an interaction with variables in the multivariate relationship, is usually statistically flawed even when the start date of the program is known. In particular, as Banerjee, Lumsdaine, and Stock (1992) point out, when the data break point is not exogenous, conventional hypothesis testing is not valid. Since in most cases the timing of introduction of the dummy variable is not in fact exogenous but is instead data-dependent, a standard "Chow" test of change in parameters is incorrect. A classic example of this problem in the time-series literature is assessment of the impact of the oil price shock. While one might have supposed, in the early 1970s, that OPEC actions would affect oil prices, the substantial impact on the macroeconomy was only

recognizable ex post. Therefore, a test of structural break in 1973 is data dependent, as the timing of the break was identified from the data.

Program introductions are often considered exogenous. However, it may not be appropriate to treat them this way when evaluating program effects. Implementation lags may make the timing of any effect endogenous since an effect can only be recognized from the data rather than an exogenously determined date. Therefore, the timing of the effect of a particular program cannot be seen as exogenous.

The time-series literature, in recognizing the data-dependence of the specification in testing for structural breaks, has taken several approaches. One relatively simple solution is to recalculate the critical values for the F-statistic when the point of break in the regression relationship is not in fact exogenous. One possibility, proposed by Hamilton (1989) and Potter (1995) is to parameterize a regime-switching mechanism and then estimate the parameters of this mechanism. While this procedure is applicable in some cases, it is not directly applicable in the program evaluation setting since the only switching is from pre-program to post-program and it is not obvious how to parameterize such a switch.

An alternative, proposed by Andrews (1993), is applied to the program evaluation setting in this paper. Rather than setting the break point a priori and testing for its statistical significance (with the corrected critical values) or parameterizing the source of the break, Andrews proposes explicitly searching for a break point. The key difference is that instead of testing for a break at a particular point in the time series, one looks over the entire time series for the location of a break point, if one exists. It is crucial to note, however, that the procedure does not necessarily find a statistically significant change in the parameters of a multivariate relationship. This feature makes the procedure well-suited for program evaluation, where the key question is whether or not the pre-program period differs from the post-program period in a statistically meaningful way.

The key advantage to this technique is that the break might not be where one thinks it is, and the procedure for testing for parameter instability makes no a priori assumptions about

location of a break, or even if one exists. Even if one knows at what point in time an intervention or program was instituted, one does not necessarily know when the program had its effect. 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. It is therefore correspondingly difficult to evaluate a program effect by testing for a change in regression parameters between two periods defined as "pre-program" and "post-program" even if one had the correct critical values for the test statistic with endogenous break point. And one may not even know precisely when the program was actually instituted (as in the example presented in Section III below).

However, with the Andrews approach, one does not need to specify a date and test for a change in parameters before and after that date. Rather, one looks for whether or not there is a break, and if so, where it. Therefore, while the technique has in practice been most prominent in the work of macroeconomists analyzing time series and structural change in parameters governing U.S. output, program evaluation is an ideal setting in which the technique can be applied with tremendous benefit.

One limitation of the technique is that it allows for only one break in the time series. In the program evaluation setting this limitation is not of great importance since in general the questions being asked are regarding the point in time when we switch from "before" to "after" (where there is assumed to be one such point), and whether the parameters of some relationship change from "before" to "after."[1] The class of problems for which the technique is applicable is quite general, however, beyond the restriction of allowing only one break. All that is required is

---

[1] It is worth brief mention, however, that a technique known as regression tree analysis (applied in Cooper 1998) can allow for multiple unknown break points in multiple dimensions. However, given the questions of interest in the program evaluation setting, the additional flexibility permitted in regression tree analysis is unnecessary. The Andrews procedure can therefore be seen as between the two extremes of assuming the location of break on the one hand, and allowing any number of breaks in any number of dimensions with the regression tree procedure on the other hand.

that the time series not follow a unit root. Beyond that, virtually any sort of "before" vs. "after" comparison one might want to do in a program evaluation setting can be incorporated.

Specifically, one can look for a change in the mean value of some outcome, as well as a change in trend. One can control for additional factors, and either test for parameter changes in the relationships between these additional controls and the outcome, or not. Thus, one can test for a change in either all the parameters of some multivariate relationship, or in any subset of the parameters. Any maximum likelihood estimation procedure is allowed.

Given a stationary time series, define a Wald statistic for the null hypothesis that the parameters of interest do not change between periods. Specifically,

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

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

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, the 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. Then one computes the Wald statistic for every possible break in the time series. Andrews then computes the asymptotic distribution for the *sup Wald* statistic, over all possible break points, and tabulates critical values.

In practice, to apply the technique in a program evaluation setting, one needs to define first 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 or some subset of the regression parameters or all

regression parameters. Then one breaks the data and computes the Wald statistic and proceeds to do this for every possible break point. Then a comparison of the maximal Wald statistic over all possible breaks to the critical value provides a straightforward test of statistical significance of program effect.

If the maximal Wald statistic exceeds the 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. If the maximal Wald statistic does not exceed the critical value then one can conclude that there was no statistically significant change in parameters. In the program evaluation setting, a finding of a maximal 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 be only indirectly attributed to the program since what has been found is a break in the parameters of a regression relationship, but one has not proven that this break is caused by the program intervention. Of course, this limitation in interpretation is also true in simple pre- vs. post- program analyses.

One parameter that must be chosen in applying this technique for searching for a data break is what Andrews refers to as "trimming." In other words, when one searches all possible locations for a break in 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 very beginning of the sample or until the very end is that there must be sufficient observations on either side of the break point under consideration to estimate the regression relationship both before and after the break point. Clearly one cannot have just one observation "pre" break or one observation "post" break.

Andrews defines a trimming parameter which specifies how far into the sample (as a percentage of the full sample size) one starts looking for a break, with a symmetric fraction of the sample left after the latest break evaluated. Note that trimming is distinct from restricting the range of observations within which one looks for a break. Limiting the window in which a break can occur is as erroneous as specifying the break point a priori since it assumes knowledge of the

location of break as exogenous. Trimming, however, can be seen as a tradeoff between being completely agnostic about the location of break (i.e., testing for breaks in as many locations as possible) and having sufficient observations before the earliest break and after the latest break to estimate the parameters.

Trimming is defined in terms of a fraction of the sample. In practical terms, with a small sample the trimming issue can be quite difficult. In small samples the inclination may be to trim a higher percentage in order to have sufficient observations to evaluate the earliest and latest break points. But at the same time one sacrifices having a large range of data points to look at. In this paper we use 15 percent trimming as a baseline, but compare these results with 10 percent and 20 percent trimming as robustness checks.

In addition, in program evaluation there is frequently not very much data after the program is assumed to have been introduced. If one trims so much as to eliminate the program introduction date from the range of possible break points the analysis may be useless from a program evaluation standpoint. Therefore, while one must remain agnostic about the existence of or location of a break in parameters, the need to trim in order to estimate parameters of interest alerts the evaluator to the importance of accumulating enough experience before testing for a break.

Within the general class of tests for parameter instability, program evaluation problems tend to focus on a particular subset of the possibilities which is different from the subset emphasized in assessing stability in the behavior of macroeconomic variables. On the other hand, program evaluations look for a change in mean rather than more complicated hypotheses. However, the procedure allows for control variables, which is often more important in program evaluation than in modeling the macroeconomy. Finally, program evaluation requires one to define the period of analysis, while macroeconomists tend to utilize the longest period for which data have been collected.

III. The Boston Gun Project

In the early 1990s, officials in many U.S. cities were concerned about the growing numbers of youth involved in homicides, both as victims and as offenders. In that environment, researchers from the Kennedy School of Government began working with the Boston Police Department to convene a working group comprising representatives of law enforcement and other agencies in the city to perform original research into Boston's youth violence problem, to craft a strategy to respond to the conditions, to implement that strategy, and to evaluate the experience. Throughout the effort, the goal of the Boston Gun Project (BGP) was to reduce youth homicide in the city of Boston in the relatively near term.

The intervention was multifaceted.[2] Members of the working group shared information to identify those involved in violent disputes and to target sanctions toward the most active individuals and groups. 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 and delivering individualized messages about the consequences of criminal activity (based on the individual's criminal history). Due to the longstanding nature of antagonisms between gangs, conflicts could be predicted with some success. In such cases, both enforcement and social services mobilized to prevent retaliations. Finally, the working group "advertised" the goals, capacities, and achievements of the initiative to individuals and gangs identified as being "at risk" of violent assault or victimization. This communication strategy was intended to deter initiations of violence and retaliations.

Two attributes of the intervention pose particular challenges for evaluation design. First, the strategy evolved over time. The working group was constituted early in 1995 and met regularly during that calendar year to conduct the research and develop the strategy. In January 1996 the group took its findings to the heads of the agencies involved for permission to move the strategy to the implementation phase. By the middle of June 1996 the working group felt the

---

[2]For the details of this effort, see Kennedy, Piehl and Braga (1996) and Piehl, Kennedy and Braga (1998).

strategy was "in place," and it was announced as such at a public meeting of the local bar association. Even once the strategy was "in place," however, the intervention's attributes changed with the circumstances of particular violent outbursts and as the nature of youth violence in the city more generally evolved. As a result, it is impossible to assert with confidence the date on which implementation of the intervention began and even more difficult to determine the date of any effect without looking at the data. Therefore the timing of any effect should be considered endogenous.

Second, much of Boston's youth violence "problem" arose out of a nexus of disputes across gangs and 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] That is, 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 the features of the BGP initiative, it 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. It is not obvious that using other cities is preferable to using control variables within Boston. A priori, there are two reasons to think that other cities are not as useful as might appear on first blush: first, trends vary greatly across cities[4] and second, there may well have been spillovers as other

---

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

[4]A multi-city examination of youth violence (in which some of the authors have been participating) show that the Boston patterns are unique among Miami, New York, Chicago, St. Louis, Pittsburgh, Atlanta, and cities in southern California.

cities adopted aspects of the program after it received positive media attention early on. In this paper we utilize the time series variation within Boston alone, applying the technique outlined above and controlling for characteristics of the city that arguably changed over time in a manner consistent with the observed changes in youth violence.

IV. Empirical Results

This section tests for a structural break in the Boston data. After describing the data, we consider the specification issues raised by this application, specifically the count nature of the dependent variable. We then present the results of this approach and compare them to the results from traditional Chow tests. Finally, we return to the issue of trimming raised above and offer a few additional robustness checks.

Data

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.[5] 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.

In Figure 1, there are two factors potentially obscuring the patterns in the data. First, the population of young people in the city fell quite dramatically over this period. Second, there is strong seasonality in the data, with August and September having the highest homicide rates. Figure 2 repeats the same series as Figure 1, after removing the month means and denominating by the population of African American males ages 15-24, as most victims are members of this

[5]Because the focus of this paper is on the application of the technique, we restrict our attention to just one outcome variable. For evidence using other outcomes, see Piehl et al. (1998).

demographic group.[6] In this figure, it is easy to see that the homicide rate is particularly low late in the time series. In the 24 months from July 1996 through June 1998, the homicide rate is at or above the average level only two times. One also sees that there is a period of lower than expected youth homicides at the beginning of the series. Even removing the month effects, substantial variation remains.

Table 1 reports descriptive statistics by year for the outcome variable, population, and several 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. (Note that 1998 does not include the high fatality months of late summer and early fall.)

There are two alternative measures for controlling for the decline in population over this period. First, the number of 18-24 year olds fell 18 percent over the period 1992-1997.[7] Second, as noted above, the vast majority of youth homicide victims come from one demographic group: young African American males. The number in this group fell by 7 percent over the time series.[8]

Table 1 includes values for several additional control variables. Indicative of the booming economy, unemployment in the city of Boston fell by over half, from 8.0 to 3.5 percent,

---

[6]Of the 155 gun and knife homicide victims aged 21 and under from 1990-1994, 88 percent were male and 78 percent were black (Kennedy et al. 1996).

[7]The population numbers are for Suffolk County rather than the city of Boston. As these entities are nearly the same (Suffolk contains Boston plus Chelsea), the changes in population are likely to be highly correlated across these geographic units. A more fundamental concern is the accuracy of year-to-year changes in population (given that an exact count is done only in Census years). Due to concerns about over-relying on the annual fluctuations, we model the homicide count as a function of the population rather than modeling the rate (which is the same as restricting the coefficient on population to be one).

[8]Our results are not sensitive to the population control used. In fact, the population variable is generally not statistically significant in the regressions reported below.

from 1992 to 1998. Violent crime also fell dramatically: robbery rates[9] (per 100,000 population) fell by 55 percent and the rate of "adult" homicide victimization[10] fell 28 percent.

Specification

As the discussion above indicated, the Andrews procedure is applicable for a broad range of different specifications. Here, we test for a change in the mean number of youth homicides. The initiative hoped to move the level of homicide to a new, lower equilibrium. 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. 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 in the constant. Before estimating the models, however, there are two remaining concerns.

First, recall that the Andrews procedure is only valid if the time series does not have a unit root. Using a Dickey-Fuller test, we reject nonstationarity in our time series (p-value = 0.0006).[11] Second, the dependent variable is counts of homicides. Because of the count nature of the data, Poisson regression is a logical choice. However, in the Poisson the variance equals the mean. As a result, a break in mean must also be a break in the variance. It is far from clear that we want to test for such a compound hypothesis. There are two reasonable alternative specifications. One could simply run OLS, correcting the standard errors for heteroskedasticity. Alternatively, given that there is a lot of skew in the data, it may be preferable to transform the dependent variable before running OLS. Taking the square root of a count variable and running

---

[9]As with population, Uniform Crime Report data are available only annually. For 1998, the preliminary figures (for the period January through June) were used.

[10]The "adult" homicide rate is defined as the number of homicide victims aged 25 and older divided by the population aged 25-44.

[11]We also ran the Dickey-Fuller test for the square root of the youth homicide count, because we rely on that specification in many of the results reported later. We handily reject a unit root there, too, with p-value = 0.0025.

OLS is recommended by Cameron and Trivedi (1998, pp.88-90). In the tables we report the results of the latter option. We note along the way the few instances in which the results are sensitive to these different specifications.

Results

Table 2 reports the results of running OLS on the square root of the monthly number of youth homicides for four sets of control variables, trimming 15 percent off of each end of the time series. All models include controls for the population of black males aged 15-24 and a full set of month dummies. The second column reports the maximum value of the test statistic for each model. For model A, the sup Wald value (32.66) occurred in June 1996. When this is compared to the Andrews critical value of 8.85 for a test of a break in one parameter with 15 percent trimming at the 5 percent significance level, we clearly reject the null of no break in mean. The effect size is a reduction of 2.45 homicide victims per month[12], or an approximately 60 percent decline. This result is not sensitive to many of the choices discussed above: changing the trim parameter to 10 percent or 20 percent, using the Poisson, and running OLS on the untransformed count all locate the maximal break in June 1996 at a level greater than the critical value.

In model B, controls for the overall unemployment rate and the robbery rate, both calculated citywide, were added to the controls for the previous model. Adding these two controls reduced the maximal value of the Wald statistic nearly in half, but the break was still placed in the same month and continued to be statistically significant. The estimated effect size was somewhat bigger with the additional controls. Neither the unemployment rate nor the robbery rate was statistically significant with the break in the maximal month.

---

[12]Note that it since it is difficult to interpret the coefficients in the square root OLS framework, this calculation comes from OLS on the levels. The measures are intended to give a rough estimate of the effect size. The estimated effect sizes are quite similar across the various specifications of the error term and dependent variable.

Model C used an alternate control variable to generate the counterfactual for youth homicide in the absence of the program. Rather than the robbery and unemployment rates, this specification used the rate of "adult" homicide victimization. Given that the intervention could very well have affected the victimization of older people (directly, if younger people reduced their victimization of older people,[13] or indirectly, because enforcement was targeted at a type of offending, not strictly on age) this is quite a strict test. With adult homicide as a control, the break continues to be located in June 1996 and the effect size is somewhat lower (2.17 victim reduction). With the break so located, the adult homicide rate has a p-value of 0.12.

Finally, model D includes all of these control variables. In this case, the maximal Wald statistic is in August 1996, with a value of 13.03. The other results are qualitatively quite similar to those discussed above. Therefore, the finding of a statistically significant decline in youth homicide in the middle of 1996 is robust to changes in control variables.

It is an interesting exercise to step back and compare inference under this procedure to the "usual" program evaluation approach in which a dummy variable turns on when the intervention is assumed to have begun. (Ignore for a moment that the evaluator might not know when that was.) The critical value for a conventional Chow test (ignoring the endogeneity of the break point) is 4.00 at the 5 percent significance level (and 7.06 at the 1 percent significance level). If one happened to have chosen the month with the maximal Wald statistic (June or August 1996 in the specifications above) one would have concluded the intervention was associated with a break. However, one would have rejected no break for many other months in the time series as well.

Figure 3 plots the test statistics from model C in Table 2.[14] The flat line (at 4.0) is the Chow (standard F-statistic) critical value. It is clear that we would have rejected no break in the

---

[13]Using data from Supplemental Homicide Reports, Cook and Laub (1998, Table 5) report that young killers 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.

[14]Because our example tests for a break in only one parameter, the Wald test statistic is equal to the F-test statistic in the Chow framework.

standard framework for the months picked out by the Andrews procedure. But, moreover, we would have rejected no break had the program dummy been placed in many of the months surrounding June 1996. Ignoring trimming, we would have rejected a Chow test if we had placed a hypothesized break in any of the following months: March 1992-January 1993, March 1993, August 1995-November 1997, and January 1998.

Methodologically, this graph points out that the endogenous break procedure makes one less likely to reject the null (the critical value is over twice the Chow critical value). It also illuminates the restriction that this approach locates only one break. The peak surrounding July 1992 represents a dramatic rise in the level of youth homicide (see Figures 1 and 2). The results in Table 2 reveal that the 1996 break was more substantial than the earlier rise. This example reflects a general concern in program evaluation: unless the world is quite stable in advance of the initiative or the econometrician can reliably model the determinants of an outcome, it can be difficult to identify important changes. As with other evaluation methods, the Andrews procedure will more reliably locate changes if the program takes place in a reasonably stable environment.

While Table 2 contains the core results of the paper, Table 3 offers some specification checks and insights into the method by altering the time frame under analysis. Because in practice one always has to decide how much data to collect, an evaluation method that is not sensitive to small changes in endpoints is preferable. For each of three different time frames, models are estimated using the three sets of controls used in Table 2. Model E covers only the period before the intervention, 1992 through 1995. Here, the maximum Wald statistic occurs early on in the period, but the test statistic does not reach the level of the critical value at the 10 percent significance level.

The other rows in the table show that changing the start and end dates a bit does not change the inference very much. The final row shows that inference about the effect of the BGP is sensitive to the controls included, as the procedure picks out a month early in the time series

(when the homicide rate was increasing) over the decrease in 1996 when all control variables are included and the time period is extended back to 1991.

As a final robustness check we varied the trimming parameter. As less of the data 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. None of our results were sensitive to whether trimming was done at 10 percent, 15 percent, or 20 percent. While trimming may not be of critical importance in much macroeconomic data, it does raise an interesting point in the program evaluation setting. The question is how much data to 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. The robustness to of the results presented here to changes in trimming suggest that the program effect is well captured with a variety of different lengths of data.

In sum, this evaluation has found that there was a statistically significant break in mean associated with substantial decreases (on the order of 60 percent) in youth homicide in the summer of 1996.[15] This time coincides with when the BGP was implemented. Controlling for the adult homicide rate we have some confidence that we have captured a program effect rather than an unrelated change in youth homicide. We are more confident in our conclusion of the existence of a program effect having used statistical methods that take into consideration the endogeneity of the break point than if we had used traditional program evaluation methods.

V. Conclusion

We have two types of conclusions: substantive and methodological. Although this paper was not intended to provide a definitive evaluation of the program, we conclude that something dramatic happened to youth homicide in the summer of 1996. Dramatic declines were

---

[15] Note that the procedure presented in this paper is more general than simply locating and testing for a break in mean. But for most program evaluations, the mean is the first order concern. If one wants to test for breaks in a larger number of parameters, the critical value increases. These values are reported in Andrews (1993) Table 1.

experienced, even controlling for adult homicide experience. Point estimates suggest the intervention was associated with a 50-60 percent decline in youth homicide. Alternative explanations (such as demographic changes, simple incapacitation, or other youth interventions) for the shift (discussed in Piehl, et al. 1998) cannot straightforwardly account for such an abrupt change.

From applying the Andrews procedure to program evaluation, we conclude that the method is flexible, easy to use, can identify timing of and test for 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.

Bannerjee, 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.

Cameron, A. Colin and Pravin K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge University Press.

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*, 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: No. 1, pp.147-183.

Piehl, Anne Morrison, David M. Kennedy, and Anthony A. Braga (1998), "Problem Solving and Youth Violence: An Evaluation of the Boston Gun Project," unpublished paper, University of California, November.

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

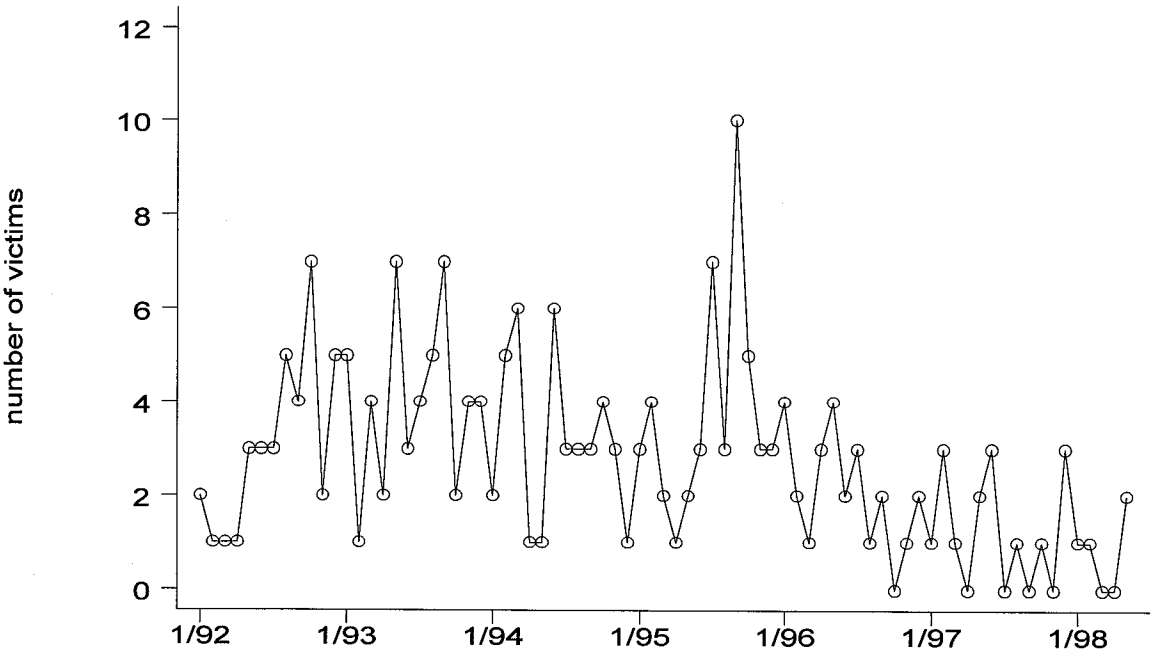

Figure 1. Monthly Youth Homicide Count 1/92-5/98

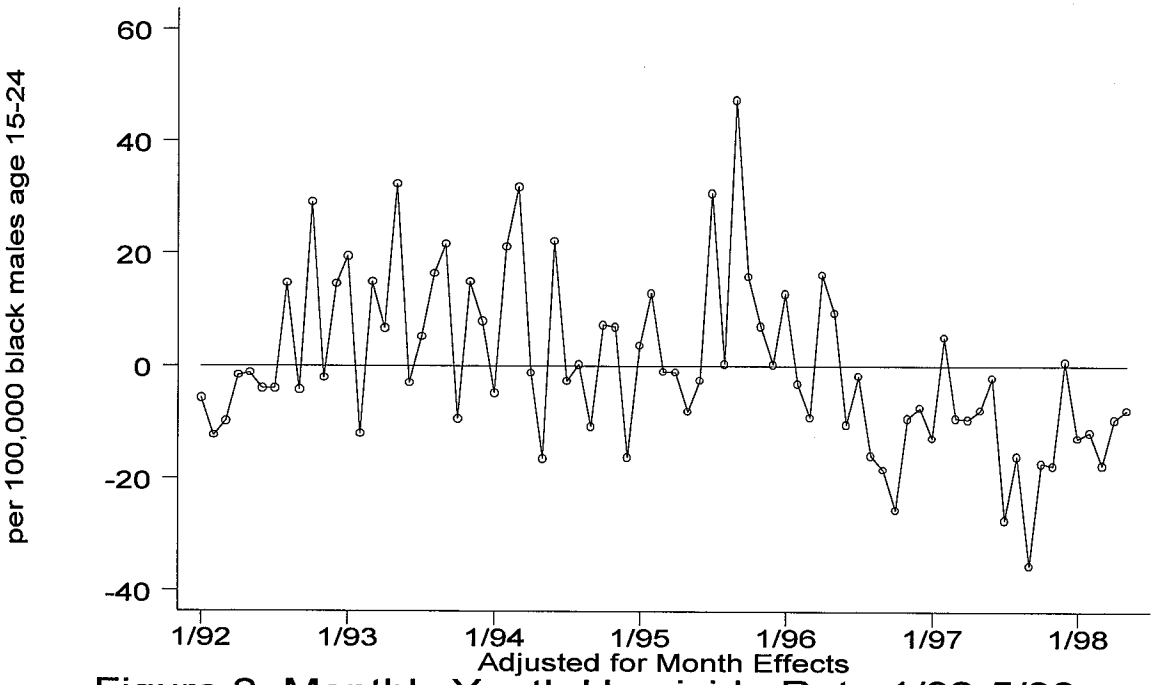

Figure 2. Monthly Youth Homicide Rate 1/92-5/98

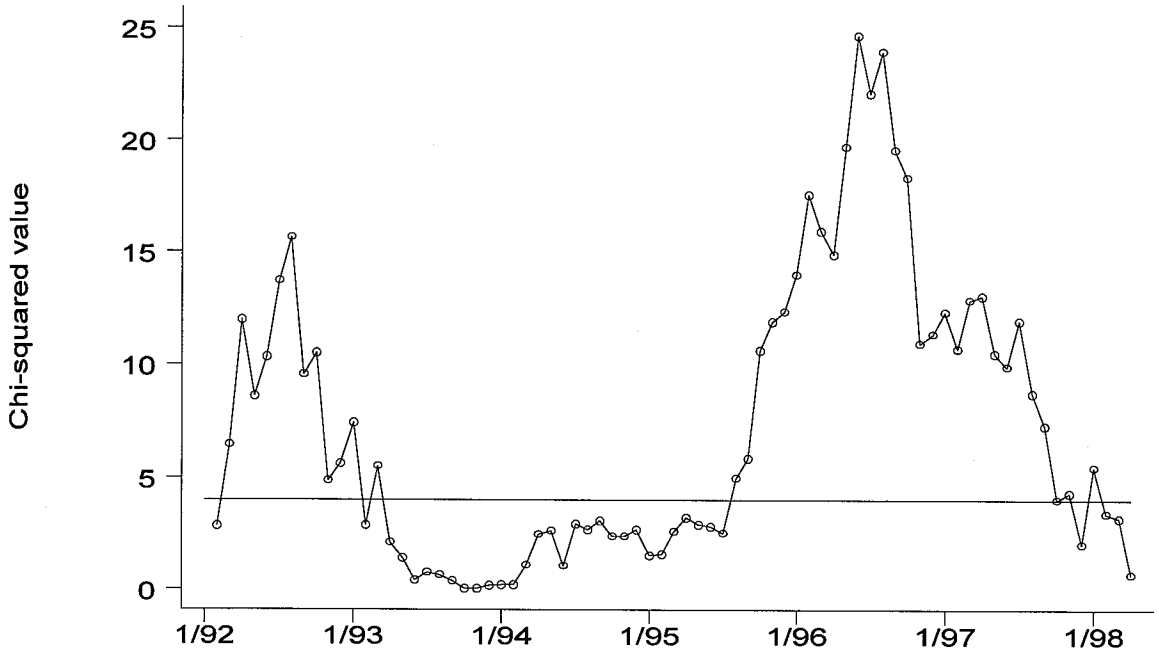

Figure 3. Test Statistic for Break in Mean

Table 1. Descriptive Statistics
(standard errors)

	Youth Homicides	Population 18-24	African American Males 15-24	Unemployment Rate	UCR Robbery Rate	Adult Homicides
1992	3.083 (0.543)	98,288	12,808	8.017 (0.153)	746	1.418 (0.227)
1993	4.000 (0.537)	94,371	12,302	6.592 (0.166)	639	1.830 (0.259)
1994	3.167 (0.520)	90,592	12,142	5.792 (0.097)	666	1.724 (0.252)
1995	3.833 (0.716)	87,399	12,078	5.300 (0.117)	557	1.803 (0.255)
1996	2.083 (0.358)	82,789	11,744	4.417 (0.124)	539	1.157 (0.205)
1997	1.250 (0.351)	80,721	11,878	4.142 (0.114)	424	1.021 (0.193)
1998	0.800 (0.374)	--	--	3.500 (0.130)	338	--

Sources: Homicide data were provided by the Boston Police Department. The other data came from various web sites: population data from the Bureau of the Census, unemployment rates for the overall labor force for the city of Boston are from the Massachusetts Department of Employment and Training, the number of robberies is from the Uniform Crime Reports collected by the Federal Bureau of Investigations (then demoninated by the overall population using Census numbers).

Notes: Only homicide and unemployment data vary by month. Others are annualized values. 1998 contains data only through May.

The adult homicide rate was calculated per 100,000 population aged 25-44.

Table 2. Parameter Instability in Youth Homicide: Breaks in Mean
Various Sets of Control Variables

Model	Max. Test Statistic	Month of Max.	Unemp. Rate	Robbery Rate	Adult Hom. Rate	Effect Size
A	32.66	June 1996	–	–	–	- 2.45 (71%)
B	17.02	June 1996	yes	yes	–	- 2.74 (79%)
C	24.62	June 1996	--	–	yes	- 2.17 (63%)
D	13.03	August 1996	yes	yes	yes	- 2.57 (75%)

Sources: See Table 1 for descriptions of the variables.

Notes: The population of black males aged 15-24 and 11 month indicators are included in all specifications in addition to the controls noted in columns (4) through (6).
N = 77 months, 1/92 through 5/98.
The Chi-squared critical values for testing for a break in a single parameter with 15 percent trimming are 7.17 at the 10% level of statistical significance, 8.85 at the 5% level, and 12.25 at the 1% level. (See Andrews 1993, Table 1.)

Table 3. Parameter Instability in Youth Homicide: Breaks in Mean
Various Time Frames

Model	Max. Test Statistic	Month of Max.	Time Period	Model B	Model C	Model D	Effect Size
E	5.16	December 1992	1/92-12/95	yes	--	--	--
E'	5.64	October 1992	1/92-12/95	--	yes	--	--
E"	5.91	October 1992	1/92-12/95	--	--	yes	--
F	8.62	August 1996	1/93-5/98	yes	--	--	-2.15 (61%)
F'	15.82	August 1996	1/93-5/98	--	yes	--	-1.70 (48%)
F"	7.34	August 1996	1/93-5/98	--	--	yes	-1.97 (56%)
G	12.96	June 1996	1/91-5/98	yes	--	--	-2.56 (72%)
G'	23.40	June 1996	1/91-5/98	--	yes	--	-2.07 (58%)
G"	11.80	August 1992	1/91-5/98	--	--	yes	+2.79 (86%)

Sources: See Table 1 for descriptions of the control variables.

Notes: The population of black males aged 15-24 and 11 month indicators are included in all specifications in addition to the controls noted in columns (4) through (6). Number of observations is indicated in column 4.
The Chi-squared critical values for testing for a break in a single parameter with 15 percent trimming are 7.17 at the 10% level of statistical significance, 8.85 at the 5% level, and 12.25 at the 1% level. (See Andrews 1993, Table 1.)