NBER WORKING PAPER SERIES

SEARCHING FOR A BREAK IN GNP

Lawrence J. Christiano

Working Paper No. 2695

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1988

Hossain Amirizadeh ably performed all the calculations in this paper. This research is part of NBER's research program in Economic Fluctuations. Any opinions expressed are those of the author not those of the National Bureau of Economic Research.

NBER Working Paper #2695
August 1988

SEARCHING FOR A BREAK IN GNP

ABSTRACT

It has been suggested that existing estimates of the long-run impact of a surprise move in income may have a substantial upward bias due to the presence of a trend break in post war U.S. GNP data. This paper shows that the statistical evidence does not warrant abandoning the no trend null hypothesis. A key part of the argument is that conventionally computed significance levels overstate the likelihood of the trend break alternative hypothesis. This is because they do not take into account that, in practice, the break date is chosen based on pre-test examination of the data.

Lawrence J. Christiano
Research Department
Federal Reserve Bank of Minneapolis
Minneapolis, Minnesota 55480

I. Introduction

There is considerable interest in measuring the long–run effects of a surprise change in income. Estimates vary widely, from a very small effect implied by a trend stationary representation for income (Deaton [1986]) to a very large effect implied by a difference stationary representation for income (Campbell and Mankiw [1988].) Ultimately, this interest stems from the view that substantive economic questions depend on the magnitude of this effect. For example, Deaton (1986) has argued that if the effect is large, then modern consumption theory is in trouble, being unable to account for the observed smoothness of consumption. Nelson and Plosser (1982) argue that the magnitude of the impact reveals the ultimate source of disturbances to the economy. If the magnitude is small, then most disturbances are to aggregate demand: for example, shocks to money, household preference, or government spending. If the effects are large, on the other hand, then most macroeconomic disturbances are supply side disturbances, such as the technology shocks emphasized by Prescott (1986).

Recently, it has been suggested that the magnitude of the impact of income innovations may have been vastly overstated, because researchers have failed to take into account that post war GNP has undergone a one–time break in trend (Perron [1986], Rappoport and Reichlin [1987].) In brief, the argument is that the change in trend is a one time innovation with permanent effect and those who ignore it confound it with the quarterly innovations, making the latter seem more long lasting than they are in fact. The trend break hypothesis has a degree of *a priori* appeal, since there are a number of "big events" of the post–war period that could have sparked a trend change. Examples are the 1964 tax cut, the oil shock of the early 1970s and financial deregulation in the 1980s.

the statistical evidence warrants abandoning the no trend break null hypothesis. I show that there is in fact little statistical evidence against the hypothesis that no break has occurred at any time in the post war period. Two difficulties make this argument less than straightforward. By elaborating on these difficulties, I hope that the results of this paper may be of use to those searching for breaks in other data series.

The first difficulty is that the standard critical values for testing the presence of a break are severely biased in favor of rejecting the no break null hypothesis. I overcome this problem by obtaining the correct small sample critical values by bootstrap methods. For example, the paper considers the case where an F statistic is used to test for a break in the intercept and slope of a trend against the null hypothesis that log GNP has a linear representation with an unbroken linear trend and two lags of log GNP. The conventional methodology in this context is to compare the computed F statistic against the 5 percent critical value of the relevant F distribution. In the application of the paper, this critical value is 3.1. I show that if the true data generating mechanism is the maximum likelihood trend stationary model with (unbroken) trend fit to post war log GNP (the TS representation), then *3.1 is in fact the 20 percent critical value* if the break being tested lies in the middle of the sample. The correct 5 percent critical value is 5.5.

The critical value discussed in the previous paragraph assumes, as does the standard one, that the break date is chosen independent of any prior information about the data being tested. This brings us to the second problem that must be confronted when searching for breaks. This arises because in practice one never selects a date to test for a break without prior information about the data. This second problem, adjusting critical values to reflect pretest examination of the data, is harder to solve than the first. The difficulty is that in practice it is hard to

translate the factors that go into selecting a particular break date into a specific algorithm that could, for example, be programmed on a computer. The paper describes several simple algorithms for selecting break dates and shows that the impact on critical values can be quite substantial, but that the impact depends sensitively on the particular break date selection algorithm used. Given the difficulty in practice of articulating precisely one's break date selection method, this sensitivity is unfortunate and complicates inference. The paper explores several options.

Perhaps the most straightforward option is to use a set of very conservative critical values which maximize the impact of pre–test data examination. Conducting inference with these critical values leaves one immune to excessive rejection of the null hypothesis of no trend break due to pre–test data examination. Unfortunately, a byproduct of its conservative nature is that the method probably has poor power characteristics. A set of critical values which is conservative in the above sense is studied in the paper. It assumes the break date was selected by choosing that break date that produces the largest F statistic for a trend break. Assuming the data are generated by the TS model, the 5 percent critical value in this case is 10.2. This is dramatically higher than the 3.1 critical value implied by the relevant F distribution. It also exceeds by far the largest F statistic in the post war GNP data, which is 6.14. In fact, the significance level of 6.14 is a little over 40 percent, relative to the TS model null hypothesis.[1] Clearly, relative to this set of critical values, there is no evidence of a trend break in the post war period. However, this set of critical values may be too conservative, entailing an

[1]For a statistic whose expected value is positive, I define its significance level as the probability, under the null hypothesis, of getting a value larger than the realized value. For a statistic with a negative expected value, I define its significance leval as the probability, under the null hypothesis, of getting a value smaller than the realized value.

unacceptable loss of power. As a result other, less conservative, critical values were computed. These also deliver no evidence to warrant rejecting the no trend break null hypothesis.

Following is an outline of the paper. The next section describes the no trend break null hypothesis of the paper. Reflecting the lack of consensus in the literature on time series models of post–war GNP, the null hypothesis is captured by two models. One represents log GNP as stationary about a trend (the TS model mentioned above) and the other represents it as first order autoregressive in first differences. These models and their fitted disturbances are used to generate the artificial data that form the basis for statistical inference in subsequent sections. Section III demonstrates the poor small sample performance of the usual F test for a trend break and supplies small sample critical values which are correct under the assumption that the choice of break date is independent of the data being studied. Section IV shows how sensitive critical values are to pre–test examination of the data. That section tabulates critical values under several alternative pre–test break date selection schemes. It shows that once pre–test and small sample distributional considerations are taken into account, the F test reveals no evidence against the null hypothesis of no trend break in post–war U.S. data. Section V considers a test for trend breaks recently introduced by Perron (1987). I show there that, although that test does not share the F test's small sample distributional problem, it is still the case that once pre–test considerations are taken into account, one cannot reject the no trend break null hypothesis. Section VI concludes the paper.

II. The No Break Null Hypothesis.

Throughout this paper the null hypothesis is that there has been no trend break in GNP. The bootstrap methodology I use requires that this null hypothesis be embedded in a completely specified time series model. Doing so is complicated by the fact that there is no professional agreement on how to model log GNP, whether as stationary about a trend (see, eg., Blanchard [1981]), or as a stationary process in first differences with no deterministic trend (Campbell and Mankiw [1988].) I avoid taking a stand on this issue by allowing for the possibility that either is correct. Accordingly, my bootstrap experiments are based on two data generating mechanisms. Each was estimated using data on log GNP covering the period 1948.1 to 1987.4. In each case, the first two quarters' observations were used up by initial conditions, leaving 158 observations for the regression. The first regression model, called the TS model, is consistently estimated under the assumption of covariance stationarity about a linear trend. Following are the results ($y_t \equiv \log GNP_t$):

TS Model

$$(II.1) \qquad y_t = \underset{(.14)}{.36} + \underset{(.00015)}{.00037}t + \underset{(.07)}{1.34}y_{t-1} - \underset{(.07)}{.38}y_{t-2} + \hat{\epsilon}_t, \quad \hat{\sigma}_\epsilon = .010.$$

Standard errors appear in parentheses. The Q statistic at lag 36 computed from the fitted residuals, $\hat{\epsilon}_t$, is 23, indicating very little evidence against the null hypothesis that there is no serial correlation in the disturbances. The $\hat{\epsilon}_t$'s are plotted in Figure 1, together with the plus and minus one standard deviation lines. The evidence in the data plot arouses no suspicion that the regression is misspecified. For example, there are no obvious outliers and the variance seems reasonably constant.

repeatedly simulating equation (1) using the actual y's for the first two quarters of 1948 as initial conditions, and obtained disturbances by randomly (with replacement) drawing from $\{\hat{\epsilon}_1,\ldots,\hat{\epsilon}_{158}\}$. By randomly drawing from the fitted disturbances in this way, I implicitly impose conditional homoscedasticity on the disturbances.[2] These 1,000 data sets are the basis for analysis of the TS model in subsequent sections.

I also estimated the following AR(1) representation for Δy_t:

DS Model

$$\Delta y_t = \underset{(.0010)}{.0050} + \underset{(.07)}{.37}\,\Delta y_{t-1} + \hat{u}_t,\ \hat{\sigma}_u = .010. \tag{II.2}$$

The Q statistic at lag 36 for this equation is 24, also indicating little evidence against the null hypothesis of no serial correlation in the disturbances. Note that, to two significant digits, the standard deviation of the $\hat{u}_t$'s are the same as that of the $\hat{\epsilon}_t$'s. In addition, the plot of the $\hat{u}_t$'s is virtually identical to that of the $\hat{\epsilon}_t$'s, and so is omitted.

I generated 1,000 data sets of y_t's, each of length 158 in the same way as was done for the TS model. In particular, the initial conditions for each simulation are the actual y_t's for the first two quarters of 1948. In addition, disturbances were obtained by randomly drawing, without replacement, from the set $\{\hat{u}_1,\ldots,\hat{u}_{158}\}$. These data sets are the basis for the analysis of the DS model that follows.[3]

---

[2]Results in Hamilton (1987) indicate there may be some room to doubt this assumption. It would be of interest to repeat the experiments in the paper with a data generating mechanism that allows for empirically plausible conditional heteroscedasticity in the disturbances.

[3]The TS and DS models are in some respects quite similar. For example, the roots of the characteristic equation of the TS model are .91 and .38, after rounding. On the other hand, the characteristic roots of the level representation for y_t implied by the DS model are 1 and .37. Because of the similarity of these roots, the effect of an

III. Critical Values That Ignore Pre–test Examination of the Data.

As a first step in looking for breaks, I estimated the following T–4 regressions:

$$(III.1) \qquad y_t = \mu + \theta d_t^i + \beta t + \gamma d_t^i t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t.$$

where,

$$(III.2) \qquad d_t^i = 0 \quad t = 1,2,\ldots,i-1$$
$$= 1 \quad t = i,i+1,\ldots,T,$$

for $i = 3,\ldots,T-2$. The estimation period is $t = 1,\ldots,T$, with $t = 0, -1$ reserved for initial conditions, and $T = 158$. The periods $t = 1$, T correspond to 1948.3 and 1987.4, respectively. The i^{th} regression allows the slope and intercept to change at date i. As such, it can accommodate both a discontinuous jump in the trend line or a continuous trend with a kink at date $t = i$. It can achieve the latter by setting $\theta = \gamma(1-i)$. Let F_i denote the F statistic for testing the null hypothesis $\theta = \gamma = 0$, ie., that there is no time trend break, in period $t = i$. The equations in (1) where used to compute $F_3,\ldots,F_{T-2}$, which are plotted in Figure 2. There I have highlighted

---

innovation on the short–term (two or three quarters) forecast horizon is very similar between the two models; however, they differ greatly in terms of the effects on the longer term forecast horizon. For example, according to the TS model a unit innovation in y_t induces the following revisions to the outlook for y_s, $s = t+1,t+2,t+3,t+4$: 1.3, 1.4, 1.4, 1.3. Thereafter the effect slowly tapers off to zero. According to the DS model, on the other hand, a unit innovation to y_t induces the following revisions to the outlook for y_s, $s = t+1,t+2,t+3,t+4$: 1.4, 1.5, 1.6, 1.6, where it remains forever. The economic consequences of these differences at long horizons can be great. For a discussion in the context of consumption theory, see Deaton (1986) or Christiano (1987).

five locally maximal F statistics. They occur on the dates 1950.1, 1965.1, 1973.2, 1980.2, and 1981.2. These F statistics are reported in Table 1, together with their significance levels, computed in a variety of ways. The significance level in the third column is based on the F distribution with 2 numerator and 152 denominator degrees of freedom.[4] Since the significance levels of the locally maximal F statistics are below 5 percent, the conventional test procedure results in a finding of a statistically significant break at each of the five dates. This can also be seen in Figure 2, where the 5 percent critical value, 3.1, of the F(2,152) distribution is plotted.

Bootstrap critical values for the F statistics where obtained by computing F statistics for dates i = 3,...,156 (=T–2) on each of the 1,000 artificial data sets generated by the TS and DS models. To discuss this further I need some notation. Denote the F statistics obtained for the i^{th} date on the r^{th} artificial data set by $F_{i,r}$, where i = 3,...,156 and r = 1,...,1000. Then, let

$$\text{(III.3)} \qquad \mathbf{F} = \begin{bmatrix} F_{3,1} & F_{3,2} & \cdots & F_{3,1000} \\ F_{4,1} & F_{4,2} & \cdots & F_{4,1000} \\ \vdots & & \ddots & \vdots \\ F_{156,1} & F_{156,2} & \cdots & F_{156,1000} \end{bmatrix}.$$

The r^{th} column of the 154×1000 matrix **F** contains the 154 F statistics computed in the r^{th} simulation. These were computed in the same way as the 154 empirical F statistics plotted in Figure 2. Two **F** matrices where computed, one based on the 1,000 artificial data sets generated by the DS model and one using the data

[4]The numerator degrees of freedom is the number of restrictions being tested. These are two: $\theta = \gamma = 0$. The denominator degrees of freedom are the number of observation in the regression minus the number of parameters in the unrestricted regression, which is 6.

generated by the TS model. I avoid making this dependence explicit in order to prevent the notation from becoming too cumbersome. The x% bootstrap critical value for F_i is the entry in row i of **F** with the property that x% of the entries in that row exceed it. Critical values corresponding to each of the TS and DS models were computed for each of i = 3,...,156, and some of these are reported in Table 2. For comparison, the bottom row of that table contains the 1, 5, 10, and 20 percent critical values of the F distribution with 2 numerator and 152 denominator degrees of freedom. Comparing the critical values of the F distribution with the bootstrap critical values shows that, with the exception of observations at the beginning and end of the data set, the bootstrap distribution is shifted to the right relative to the F distribution. Moreover, the bootstrap critical values associated with the DS model are shifted to the right of those associated with the TS model. These shifts are also evident in Figure 2, which plots the 5 percent critical values implied by the TS and DS model. The fact that the bootstrap distribution is shifted to the right relative to the F(2,152) distribution implies that using the F distribution will result in too many rejections of the null hypothesis of no trend break.

The simulated F statistics allow me to compute bootstrap significance levels for the empirical F statistics reported in Table 1. The columns marked TS and DS report the significance levels assuming the data are generated by the TS and DS models, respectively. Note that the F statistics corresponding to all dates after 1965.1 fail to be significant even at the 10 percent level relative to the bootstrap F distributions. The 1965.1 break is not significant at the 5 percent level under either distribution, and it is not significant even at the 20 percent level assuming the data are stationary about trend. There is more evidence of a break in 1950.1. According to the distribution of the F statistic implied by the TS model, the break is significant in 1950.1 at even the one percent level. It is somewhat less significant if

one assumes the data have a unit root, though one would still reject the null hypothesis of no trend break at the 5 percent level. These observations can also be made by inspecting Figure 2. Figure 3 plots the log GNP data used in the study and also a time trend with break in 1950.1.[5]

In sum, this section documented that the critical values from the F distribution are far too small for testing for a break in U.S. GNP data, even if one ignores the fact that break points being tested were determined by pretest examination of the data. The point is dramatized in Figure 2. It shows that the 5 percent bootstrap critical values lie far above that implied by the relevant F distribution. The shift is sufficiently large that many trend breaks that look statistically significant relative to the F distribution are, in fact, not. The only break points that still look as though they may be statistically significant are those occurring in the early 1950s.

[5]The trend line in Figure 3 was based on estimating (III.1) by ordinary least squares with $\phi_1 = \phi_2 = 0$ and i corresponding to 1950.1.

IV. Taking Account of Pre–Test Examination of the Data.

Let B denote the date on which a trend break occurs under the alternative hypothesis of a test. The sampling results in the previous section assumed that B was determined independent of the data being tested. This section shows that plausible ways of endogenizing the choice of B result in higher critical values for the F test for a trend break. In particular, what little evidence remained in the previous section that there is a trend break in GNP disappears completely once endogeneity of B is taken into account.

The first subsection formally defines six ways of endogenizing B. The fact that these models of B are explicit mathematical functions of the data reflects the requirements of my analysis. It does not reflect a view that investigators necessarily use mathematical formulas to determine B in practice. The hope is that the mathematical algorithms studied approximate reasonably well the more informal process of selecting B that investigators actually use. In many cases, they choose B based on a visual examination of the data, or of some related series, or based on the suggestion of others who have done so.

After presenting the break date selection algorithms, I report the size of the F test for a structural break when B is in fact endogenous, but the critical values discussed in the previous section—which ignore the endogeneity of B—are used to conduct inference.[6] There I show that if one applies conventional testing practice, one can find a trend break in almost all realizations from the TS and DS models. In the third subsection I report pre–test adjusted critical values for the F statistic. There I report maximal critical values which have the property that if they indicate

---

[6]The size of a test is the probability of falsely rejecting the null hypothesis.

rejecting the null hypothesis of no trend break, then one can do so without fear of that the probability of a false rejection is unduly high.

IV.a Six Break Date Selection Methods Defined

The first three of the six algorithms for choosing B selects the maximal F statistic from a subset of dates in the sample. Each is a special case of what I call the F Max method. The first of these, called the F Max_{untr} method, selects the maximal F statistics from the untruncated set $\{F_3,...,F_{156}\}$. As is clear from Figure 2, the empirical break date chosen by F Max_{untr} is 1950.1. Inspection of Figure 3 suggests that it is not at all implausible that an investigator examining the post war GNP data might conjecture that a large jump in the economy's trend might have occurred around that date. However, this is not the only plausible mechanism for endogenizing the selection of the break date, since others have investigated the hypothesis that a break occurred at other dates, such as 1973.2. Evidently, F Max_{untr} does not approximate well the method used by these investigators. Following are two simple algorithms that can account for the choice of 1973.2 as a date to test for a trend break. The first captures a suspicion felt by many that a trend break occurred in the early 1970's as a consequence of the first oil shock. That algorithm, called F Max_{oil}, identifies the break date with the date on which max $\{F_{99},...,F_{122}\}$ occurs. This corresponds to the interval of time 1973.1 to 1978.4, inclusive. Another break date selection method, F Max_{1970s}, selects the date on which max $\{F_{87},...,F_{126}\}$ occurs. The dates t = 87 to 126 correspond to the interval 1970.1 to 1979.4. The rationale for F Max_{1970s} is that it reflects the sense that there has been a "productivity slowdown" in the 1970's, whose exact date is unknown. Evidently, both F Max_{oil} and F Max_{1970s} choose 1973.2 as the empirical

break date. An advantage of studying F Max_{untr}, F Max_{oil}, and F Max_{1970s} is that it permits judging the sensitivity of critical values to the width of the interval over which one looks for a maximal F statistic.

The three other break date selection algorithms are motivated by the observation, evident in Figure 2, that under the null hypothesis of no trend break, F statistics for different B's are realizations from different distributions. In particular, the distribution of F's corresponding to B's in the middle of the data set are shifted to the right relative to F's at the beginning and end. This suggests that a more plausible break date selection algorithm is not to choose the date of the maximal F statistic, but instead the date on which the F statistic with smallest significance level occurs. Thus, if an F statistic early in the data set is smaller than one somewhere in the middle, it might make more sense to select the former as the most likely break date if its significance level under the null hypothesis of no break is smaller. For example, according to Table 1 the F statistic for B = 1973.2 exceeds that for B = 1979.4, although the significance level of the latter is less than that of the former.

Three minimum significance level techniques for selecting B where chosen by analogy with the three discussed in the previous paragraph. In particular, the Min Sig_{untr} method selects the date from the period 1949.1 to 1987.2 with the F statistic having the smallest significance level. Similarly, Min Sig_{oil} limits the break date to occurring in the period 1973.1 – 1978.4 and Min Sig_{1970s} limits it to the period 1970.1 – 1979.4. Since the significance level is a function of the model of the null hypothesis, there is a set of Min Sig methods corresponding to the TS model and one corresponding to the DS model.

When the untruncated sample is considered, both the DS and TS versions of the Min Sig_{untr} method select 1950.1 as the break. The DS version of Min Sig_{oil}

chooses 1978.3 as the most likely date of the break, while the TS version selects 1973.2. Finally, both the DS and TS versions of Min Sig_{1970s} select 1979.4 as the most likely date of the break. I do not make the notation for Min Sig explicitly reflect whether it is based on the DS or TS model in order to avoid proliferating symbols.

IV.b Impact on Size of Endogenizing the Choice of Break Date.

Table 3 reports the size of F tests for trend breaks which ignore pre–test data examination when in fact one of the six break date selection methods introduced in the last section are used. There are three panels in Table 3, each of which corresponds to a different subset of dates from which break dates were picked. There are seven columns. Columns 2 – 5 pertain to the F Max method, whereas columns 6 and 7 pertain to the Min Sig method. Numbers in italics are results based on the TS model, and the numbers in columns 2, 4, and 6 are based on the DS model. Each number in columns 2 and 3 is the fraction of times out of 1,000 that the F Max F statistic exceeds the critical value of indicated size from the F(2,152) distribution. Results in columns 4 and 5 are the fraction of times that the bootstrap critical values discussed in section III are exceeded. Columns 6 and 7 report the fraction of times that the minimum significance level is below the corresponding significance level in column 1.

The most dramatic results appear in the first column of Panel A in Table 3. This shows that when the data are generated by the TS model, the break date is selected by the F Max_{untr} method, and the conventional practice of using critical values from the F distribution is followed, then a test with nominal 5 percent size *in fact has size 98 percent.* When the data are generated by the DS model, then the

size of this test is 100 percent, after rounding. Of course, this enormous frequency of rejections reflects in part the fact—demonstrated in section III—that the critical values of the F distribution are far too small, even when B is exogenous. Once this is taken into account, then the size of the F test falls, as is indicated in columns 4 and 5. Nevertheless, the size continues to be extremely large. For example the size of the F test based on a break date selected by the F Max_{untr} method, which uses the pre–test unadjusted bootstrap critical values with 5 percent nominal size, in fact has size 65 percent. Looking at panels B and C, we see that as the interval of dates from which the break date is selected shrinks, the size of the pre–test unadjusted test falls. However, even when only the six year period 1973.1 – 1978.4 is considered, the size of the nominal 5 percent test is still around 20 percent. Of course, in the limit as the interval of dates shrinks to unity, the size of the tests based on the bootstrap critical values converge by construction to the nominal size of the test. The size of the pre–test unadjusted F test when the Min Sig method of selecting break dates is used is roughly the same as the size of the F Max method. Where the two differ, it is always the size of Min Sig that is the larger, by construction.

IV.c Critical Values of Pre–Test Adjusted Tests for Trend Break

Critical values of the F test for trend break which reflect the several ways of endogenizing B discussed in subsection IV.a are reported in Table 4. Not surprisingly, it takes a much larger F statistic to reject the null hypothesis $\theta = \gamma = 0$ when the break date has been selected as a function of the data than when it has not. For example, when the break date is selected by the F Max_{untr} method and the null hypothesis is the TS model, then it takes an F statistic of 9.0 to reject the

null hypothesis at the 10 percent level. This is to be compared with the 4.3 critical value that applies if the break date is selected exogenously to be in the middle of the sample. Alternatively, if the break date is selected by Min Sig_{untr} then the F statistic has to have an unadjusted significance level of .2 percent to be significant at the 10 percent level. The critical values in the Table can be used to assess the significance of the empirical F statistics reported in Table 1.

Consider first the possibility of a break in 1950.1. This date was chosen as the most likely break date by both F Max_{untr} and Min Sig_{untr}. Relative to the DS model, a maximal F statistic of 6.14 is actually quite *small*, having a significance level around 70 percent (see panel A, Table 4.) This contrasts sharply with the 1.1 percent significance level implied by the pre–test unadjusted bootstrap critical value and the .3 percent significance level implied by the F distribution (see Table 1.) Similarly a maximal F statistic of 6.14 has a significance level in excess of 40 percent relative to the TS distribution. Thus, in fact there is no basis for rejecting the null hypothesis of a trend break in 1950.1. This stands in striking contrast with the implications of the conventional testing methodology, which would result in a finding that the evidence of a trend break is considerable. This result is dramatized in Figure 2, which shows how much the pre–test adjusted 5 percent critical values exceed all the empirical F statistics.

There is also no basis for rejecting the null hypothesis of a break at any of the other dates listed in Table 1 at the conventional 5 percent level. This is true even before taking pre–test data examination into account. When this is done, the significance levels of the test statistic rise, making it only harder to reject the null hypothesis. It is, nevertheless, instructive to investigate the case for a break in 1973.2.

The evidence pertaining to the possibility of a trend break having occurred in

1973.2 is collected in Table 5. That table illustrates the central points of this paper. The first three columns are taken from Table 1 for ease of comparison. As noted previously, they illustrate how much the F distribution understates the significance level of a trend break. The right four columns show how failure to take pre–test examination of the data into account also results in understating the significance levels. If the 1973.2 break date had been selected from the oil shock period, then the significance level is between 40 and 60 percent, depending on whether one interprets the F statistic relative to the TS or the DS model. If instead the 1973.2 break date was selected after inspecting all dates in the 1970s, then the significance level jumps even further, to the 60 to 75 percent range. Besides showing that there is absolutely no statistical evidence against the null hypothesis $\theta = \gamma = 0$ in 1973.2, the results illustrate the difficulty of assigning a precise pre–test adjusted significance to an F statistic. A researcher may report that he/she selected the 1973.2 date by examining a very limited set of dates only, however there might still be room to wonder whether the researcher's choice was influenced by the advice of someone else who suggested looking for a break in 1973 based on examining *all* the data in the 1970s, or even more data. As Table 5 shows, whether or not the latter is true has a quantitatively large impact on the significance level of the F statistic. Although in the present case this impact does not affect the outcome—that there is no evidence of a break—one can imagine other cases where it does. This problem is avoided if the investigator were required to present a break date selection algorithm which selects his/her break date as a function of *all* the observations (eg., F Max_{untr}), and perhaps observations on some other data series as well. In the present case, those who argue for a single trend break in 1973.2 would have to argue why *that* is the most likely date for a break and not, for example, 1950.1 or 1965.1. The researcher's break date selection algorithm could then be used to compute the

significance level of the test statistic for a break.

V. Perron's Modified Dickey–Fuller Test For a Trend Break.

Perron (1987) and Rappoport and Reichlin (1987) argue that the failure of the Dickey–Fuller test to reject the unit root hypothesis reflects not the presence of the unit root, but instead that the data are trend stationary about a broken trend. Perron (1987) proposes a modification to the Dickey–Fuller test which permits, under the alternative to the unit root null hypothesis, that the data are stationary about a broken trend. He tabulates a set of critical values for his test statistic which assume that the break date is picked exogenously. When these critical values are used to interpret his test statistic computed using post war U.S. GNP data, the unit root hypothesis is rejected against the broken trend alternative at the 10 percent level. This section shows that when the critical values are pre–test adjusted, then the unit root hypothesis can be rejected at only the 15 to 30 percent level, depending on the exact break date selection algorithm used. Thus, like the F test of the previous sections, the Perron statistic offers no evidence of a trend break in post war GNP.[7]

In the context of postwar, quarterly U.S. GNP, Perron proposes estimating the following augmented Dickey–Fuller regression:

$$\text{(V.1)} \qquad \Delta y_t = \mu + \theta d_t^i + \beta t + \gamma d_t^i \times t + \alpha y_{t-1} + c_1 \Delta y_{t-1} + c_2 \Delta y_{t-2},$$

where d_t^i is defined in (II.2).[8] Under the null hypothesis of Perron's modified

---

[7]Perron applies his test to many other data series, and reports evidence of trend breaks. It may well be that those results are robust to date break selection considerations. It would be of interest to investigate this.

[8]In some contexts, Perron advises imposing the restriction $\theta = \gamma(1-i)$. In this case, the alternative hypothesis is a model with continuous trend, but a possible change in slope at date i.

Dickey–Fuller test, $\theta = \gamma = \alpha = \beta = 0$. In this case, (1) is a difference stationary model which, when $c_2 = 0$, reduces to the DS model of this paper. Perron recommends comparing the t statistic on α, t_α, with the critical values tabulated in his paper. In the light of the analysis of the F test in section II, it is not surprising that those critical values depend on the date, B, on which the trend break is permitted to occur under the alternative hypothesis. Asymptotic critical values for t_α relevant when the break date is exogenously set at 1973.2 are reported in the first row of Table 6. These are taken from Perron. The second row of Table 6 reports bootstrap critical values for t_α computed using the 1,000 data sets generated by the DS model described in section II. Note the similarity of these two sets of critical values. Thus, unlike the F statistic, the Perron statistic has roughly the same sampling distribution in a sample the length of post war quarterly data as it does asymptotically. Subsequent rows in Table 6 provide pre–test adjusted critical values for a variety of break date selection algorithms. The first three are the three F Max algorithms discussed in the previous section. Results for the Min Sig algorithm are not reported since they are roughly the same as those for the F Max method. The last three rows in Table 6 report results based on the Min t_α break date selection method. It selects break dates to minimize t_α over the indicated set of dates.

Table 6 indicates that the effect of selecting the break date as a function of the data being tested is to increase the likelihood that Perron's version of the Dickey–Fuller regression will spuriously reject the unit root specification in favor of the broken trend alternative. This is because the pre–test adjusted critical values exceed the unadjusted critical values.[9] Note that the critical values implied by the

[9]The pre–test unadjusted critical values in Table 6 assume a break date near the middle of the sample. These critical values are larger than those closer to the beginning or the end of the sample (see Table 5B in Perron). Consequently, the

Min t_α break date selection procedure are roughly the same as those implied by F Max. Where there is a difference, the Min t_α critical values are larger, by construction.

The column marked t_α Table 7 reports empirical values of t_α and associated significance levels, under alternative break date selection mechanisms. The numbers in braces in that column report the expected value of t_α. Expected values and significance levels were computed using the 1,000 data sets generated by the DS model. The break date preferred by Perron is 1973.2. The first row of Table 7 reports that t_α for that date is –3.94, and that the probability is only 8 percent of getting a value of t_α smaller than –3.94 in the 100th date (1973.2) of a sample generated by the DS model. The expected value of t_α is –2.81, and this evidently is quite far from the empirical value of –3.94. Thus, the null hypothesis is rejected at the 9 percent significance level, assuming 1973.2 is picked exogenously.

Subsequent rows in Table 7 report pre–test adjusted results. Consider first the results corresponding to F Max_{1970s} and F Max_{oil}. As noted before, these two break date selection procedures rationalize picking 1973.2 as a break date. However they assign very difference significance levels to the empirical estimate of t_α = –3.94. For example, F Max_{1970s} assigns it a significance level of 23 percent. If we evaluate t_α = –3.94 relative to this break date selection procedure, then the null hypothesis cannot be rejected at even the 20 percent significance level. Not surprisingly, the significance level drops if t_α = –3.94 is evaluated relative to F Max_{oil}. In that case one can reject the null hypothesis at the 20 percent level, although not at the 15 percent level. The two untruncated break date selection algorithms pick different dates. According to Min $t_{\alpha^{untr}}$ the most likely break date

---

pre–test adjusted critical values in Table 6 exceed these other unadjusted critical values by even more.

is 1965.1, whereas F Max_{untr} identifies 1950.1 as the most like break date. The associated significance levels are 18 and 34 percent. Here too, there is no evidence of a statistically significant break.

The last column in Table 7 reports the results of a chi–square test of the null hypothesis $\theta = \beta = \gamma = \alpha = 0$. The chi square statistic is 16.99 when B is set to 1973.2. The first row reports that the significance level of 16.99 is 11 percent when the break date is treated as though it had been selected exogenously. When instead it is interpreted relative to the F Max selection procedure, 16.99 has a significance level of 27 or 39 percent. In fact, the null hypothesis cannot be rejected at even the 20 percent level relative to any of the break date procedures. Evidently, this test also delivers no evidence of a trend break in post war U.S. GNP.

VI. Conclusion

This paper has tested the null hypothesis that the parameters of the time series model for GNP have been stable during the post war period against the alternative that there has been a one time break in trend. A variety of test statistics were presented and none can reject the null hypothesis at even the 15 percent level. In reaching this conclusion a pitfall that confronts tests for structural break was identified, and a bootstrap simulation methodology for overcoming it applied. The problem arises because the standard sampling theory used to interpret tests for structural break assumes, implausibly, that the date of the break is chosen independent of prior information about the data, or some related series. In practice researchers use a combination of visual examination of data plots, consultation with colleagues, and formal techniques to select a break date which is then tested for statistical significance. I showed that whether or not the computed statistical significance level takes into account pre–test examination of the data can make a drastic difference. For example, the F statistic of the null hypothesis that a trend break occurred in 1950.1 is 6.14. This statistic has significance level .5 percent under the null hypothesis that post–war log GNP data are stationary about an unbroken linear trend and assuming that the date 1950.1 was chosen exogenously. With such a low significance level one would ordinarily reject the no break null hypothesis easily. However, a critical assumption underlying such an inference fails The date 1950.1 was not picked at random. Instead it was chosen because it is the date in the sample which produces the largest F statistic testing for a break. When this is properly taken into account, the significance level of the 6.14 F statistic jumps from .5 percent to over 40 percent.

Table 1: Selected Empirical F Statistics and their Significance Levels

Break Date	F Statistic[1]	Pre-test Unadjusted Significance Level[2] F(2,152)[3]	*TS*[4]	DS[5]
1950.1‡	6.14	.003	*.005*	.011
1965.1‡	4.44	.013	*.083*	.233
1973.2‡	3.17	.045	*.162*	.348
1978.3	2.74	.068	*.179*	.304
1979.4	3.06	.050	*.134*	.225
1980.2‡	3.41	.036	*.121*	.189
1981.2‡	3.15	.046	*.127*	.183

[1]Empirical F statistics testing for a trend break in the period indicated in column 1.
[2]Significance level assuming break date is selected without prior examination of the data being tested.
[3]Significance level of the associated column 2 F statistic using the F distribution with 2 numerator and 152 denominator degrees of freedom.
[4]Significance level of the associated column 2 F statistic assuming the data are generated by the TS model described in section II.
[5]Significance level of the associated column 2 F statistic assuming the data are generated by the DS model described in section II.
‡These dates are highlighted in Figure 2.

Table 2: Critical Values of F Statistic[1]

t/158	date	DS Model 1%	5%	10%	20%	*TS Model* 1%	5%	10%	20%
.02	49,1	6.3	3.5	2.5	1.6	*5.4*	*3.6*	*2.7*	*1.6*
.10	52,2	6.5	4.3	3.4	2.5	*5.8*	*4.0*	*3.3*	*2.2*
.20	56,2	7.7	5.9	4.8	3.4	*6.9*	*4.9*	*3.9*	*2.8*
.30	60,1	9.3	6.4	5.1	4.1	*7.2*	*5.5*	*4.1*	*2.9*
.40	64,1	10.0	7.1	6.0	4.8	*7.5*	*5.4*	*4.3*	*3.0*
.50	68,1	10.6	7.7	6.1	4.8	*7.7*	*5.5*	*4.3*	*3.2*
.60	72,1	10.0	7.5	6.1	4.8	*8.1*	*5.0*	*4.1*	*3.0*
.70	76,1	9.4	6.7	5.4	3.9	*6.4*	*4.7*	*3.9*	*2.8*
.80	79,4	8.6	5.9	4.6	3.2	*6.9*	*4.6*	*3.6*	*2.5*
.90	83,4	7.2	4.7	3.6	2.3	*6.3*	*3.9*	*3.0*	*2.0*
.99	87,1	5.6	3.1	2.3	1.6	*5.3*	*3.2*	*2.3*	*1.6*
F(2,152)		4.8	3.1	2.3	1.6				

[1]Rows 1 – 11 provide the critical values, for the indicated set of dates (year,quarter), size and data generating mechanism, of simulated F statistics. The last row provides the critical values of the F distribution with 2 numerator and 152 denominator degrees of freedom.

Table 3: Size of Pretest Unadjusted Trend Break Tests When Break Dates are Selected Endogenously and the Data Are Generated by the DS Model and *TS Model*[1]

Nominal Size[2]	F Max Method				Min Sig Method[5]	
	F(2,152) Critical Values[3]		Bootstrap Critical Values[4]			
	A. Untruncated Break Date Selection Methods[6]					
1%	.91	*.73*	.27	*.29*	.30	*.31*
5%	1.00	*.98*	.65	*.72*	.76	*.76*
10%	1.00	*1.00*	.83	*.91*	.92	*.94*
20%	1.00	*1.00*	.95	*.99*	.99	*.99*
	B. Productivity Slowdown Break Date (1970s)[7]					
1%	.50	*.31*	.07	*.09*	.08	*.09*
5%	.76	*.60*	.24	*.28*	.25	*.29*
10%	.86	*.76*	.39	*.45*	.40	*.45*
20%	.95	*.89*	.59	*.65*	.61	*.66*
	C. Oil Shock Break Date[8]					
1%	.38	*.21*	.05	*.06*	.05	*.06*
5%	.63	*.45*	.17	*.20*	.17	*.20*
10%	.74	*.60*	.29	*.32*	.30	*.32*
20%	.87	*.77*	.47	*.52*	.48	*.52*

[1]Frequency of times, out of 1000, that the null hypothesis, $\theta = \gamma = 0$ is rejected when i in equation (III.1) is chosen by one of the six methods described in subsection IV.a. The *italicized* numbers are the results obtained when the TS model was the data generating mechanism, the other results are based on the DS model.

[2]Size of critical value used to assess the presence of a break. This size ignores that the break date itself was chosen as a function of the data prior to executing the test.

[3]Presence of a break is tested by comparing F Max with the critical values of the F(2,152) distribution, which are reported in the bottom row of Table 1.

[4]F Max is compared with the bootstrap critical values for the F statistic discussed in section III, and reported for selected dates in Table 1.

[5]Frequency of times that the minimum significance level is below the indicated nominal significance level.

[6]Untruncated break date selection methods consider the possibility of a break in dates 3,...,156.

[7]Break dates chosen from the restricted interval t = 87,...,126, which corresponds to 1970.1 – 1979.4.

[8]Break dates chosen from the restricted interval t = 99,...,122, which corresponds to 1973.1 – 1978.4.

Table 4: Critical Values of Pre-Test Adjusted Tests for Structural Break DS Model and *TS Model*[1]

Break Date Selection Method	1%	5%	10%	20%	70%	80%	90%	99%
A. Data Generating Mechanism: DS Model								
F Max[2]								
Untruncated	15.5	12.7	11.3	9.7	6.1	5.6	4.8	3.4
1970's	14.6	10.5	9.1	7.4	3.5	2.8	2.1	1.0
Oil Shock	13.0	9.6	8.2	6.3	2.6	2.0	1.4	0.5
Min Sig[3]								
Untruncated	0.0	0.1	0.2	0.5	4.2	5.8	8.8	18.4
1970's	0.1	0.5	1.3	3.6	27.3	36.4	47.7	71.3
Oil Shock	0.1	1.0	2.3	6.0	37.9	47.9	61.3	82.4
B. Data Generating Mechanism: *TS* Model								
F Max								
Untruncated	*12.8*	*10.2*	*9.0*	*7.8*	*4.9*	*4.3*	*3.8*	*2.7*
1970's	*10.8*	*8.2*	*7.1*	*5.8*	*2.6*	*2.1*	*1.5*	*0.7*
Oil Shock	*10.4*	*7.3*	*6.0*	*4.9*	*1.9*	*1.4*	*1.0*	*0.5*
Min Sig								
Untruncated	*0.0*	*0.0*	*0.2*	*0.5*	*4.0*	*5.7*	*8.0*	*15.9*
1970's	*0.0*	*0.5*	*1.1*	*2.7*	*22.0*	*30.3*	*42.3*	*66.3*
Oil Shock	*0.1*	*0.8*	*1.8*	*5.0*	*33.0*	*44.0*	*57.2*	*76.8*

[1]x is a y% critical value if the probability (here defined as frequency, out of 1000 trials) of exceeding x is y%, under the null hypothesis. In panel A the null hypothesis is the DS model, and in panel B it is the TS model.

[2]The critical values for F Max were obtained as follows. First in each of the DS and *TS* cases, the 1000 simulated F Max statistics were ranked, with the smallest one ranked 1 and the largest ranked 1000. The F Max statistic with rank 990 is the 1% critical value, the one with rank 950 is the 5% critical value, and so on.

[3]The critical values for Min Sig were obtained in the same way as for the F Max critical values. That is, in each of the DS and *TS* cases the 1000 simulated Min Sig statistics were ranked, with the smallest one ranked 1 and the largest ranked 1000. The Min Sig statistic with rank 990 is the 1% critical value.

Table 5: Significance Level of 3.17 F Statistic Testing a Trend Break in 1973.2

Pre-test Unadjusted[1]			Pre-test Adjusted[2]			
			F Max_{oil}		F Max_{1970s}	
F(2,152)	*TS*	DS	*TS*	DS	*TS*	DS
.045	*.162*	.348	*.430*	.607	*.584*	.746

[1]Entries in these columns are taken from Table 1.
[2]Entries under F Max_{oil} are the fraction of times, out of 1,000, that the F Max_{oil} statistic exceeded the empirical F statistic. This was computed relative to the artificial data generated by the TS and DS models, as indicated. The entries under F Max_{1970s} were obtained in a similar way, based on the simulated F Max_{1970s} statistics.

Table 6: Critical Values of t_α[1]

Break Date Selection Method	1%	5%	10%	20%
1973.2(asymptotic)[2]	-4.84	-4.22	-3.92	
1973.2	-4.91	-4.23	-3.88	-3.51
F Max				
untruncated	-5.86	-5.23	-4.86	-4.45
1970's	-5.34	-4.79	-4.42	-4.00
oil shock	-5.18	-4.62	-4.25	-3.81
Min t_α				
untruncated	-5.86	-5.23	-4.93	-4.55
1970's	-5.38	-4.82	-4.46	-4.10
oil shock	-5.18	-4.66	-4.28	-3.88

[1]With the exception of the results in the first row, the data generating mechanism underlying the results in this table is the DS model. Critical values are for the t statistic on α in (V.1) with k = 2.
[2]Computed by interpolating the relevant entries in the $\lambda = 0.6$ and $\lambda = 0.7$ columns in Perron (1987,Table 5B). Weights of .7 and .3 were assigned to the $\lambda = 0.6$ and $\lambda = 0.7$ columns, reflecting that 1973.2 roughly corresponds to a value of $\lambda = 0.63$. Here, λ is the ratio of the break date to the number of observations in the sample.

Table 7: Point Estimates, (Significance Levels), and {Expected Values} of Two t Statistics Based on the Following Regression Equation:[1]

$$\Delta y_t = \mu + \theta d_t^B + \beta t + \gamma d_t^B \times t + \alpha y_{t-1} + c_1 \Delta y_{t-1} + c_2 \Delta y_{t-2}.$$

$$d_t^B = \begin{cases} 0 & t < B \\ 1 & t \geq B \end{cases}$$

Break Date Selection Method	Empirical Break Date, B	t_α	λ [2]
1973.2	1973.2	-3.94 (.08) {-2.81}	16.99 (.11) {11.57}
F Max			
Untruncated	1950.1	-4.09 (.34) {-3.70}	20.29 (.46) {20.50}
1970s	1973.2	-3.94 (.23) {-3.28}	16.99 (.39) {16.20}
oil shock	1973.2	-3.94 (.16) {-3.06}	16.99 (.27) {14.51}
Min t_α			
Untruncated	1965.1	-4.63 (.18) {-4.01}	20.54 (.36) {19.28}
1970s	1974.1	-3.96 (.25) {-3.43}	16.23 (.40) {15.48}
oil shock	1974.1	-3.96 (.18) {-3.20}	16.23 (.28) {13.88}

[1]The data generating mechanism for expected values and significance levels is the DS model.

[2]λ is the likelihood ratio statistic for testing the null hypothesis $\theta = \beta = \gamma = \alpha = 0$. It is computed as $\lambda = (T-c)(\sigma_r^2 - \sigma_u^2)$, where T is the number of observations in the sample (T = 157), c (=7) is a correction for small sample bias (see Sims[1980,p.17]), σ_r^2 and σ_u^2 are the sum of squared errors in the restricted and unrestricted regressions, respectively.

References

Blanchard, Olivier J., 1981, "What is Left of the Multiplier Accelerator?", American Economic Review Papers and Proceedings vol. 71, no. 2, pp.150–4, May.

Campbell, John and N. Gregory Mankiw, 1987, "Are Output Fluctuations Transitory?", Quarterly Journal of Economics, November.

Christiano, Lawrence J., 1987, "Why Is Consumption Less Volatile Than Income?", Federal Reserve Bank of Minneapolis Quarterly Review, Fall.

Deaton, Angus, 1986, "Life–cycle Models of Consumption: Is the Evidence Consistent With the Theory?" National Bureau of Economic Research Working Paper 1910.

Hamilton, James D., 1987, "A New Approach To The Economic Analysis of Nonstationary Time Series and the Business Cycle," manuscript, University of Virginia, July.

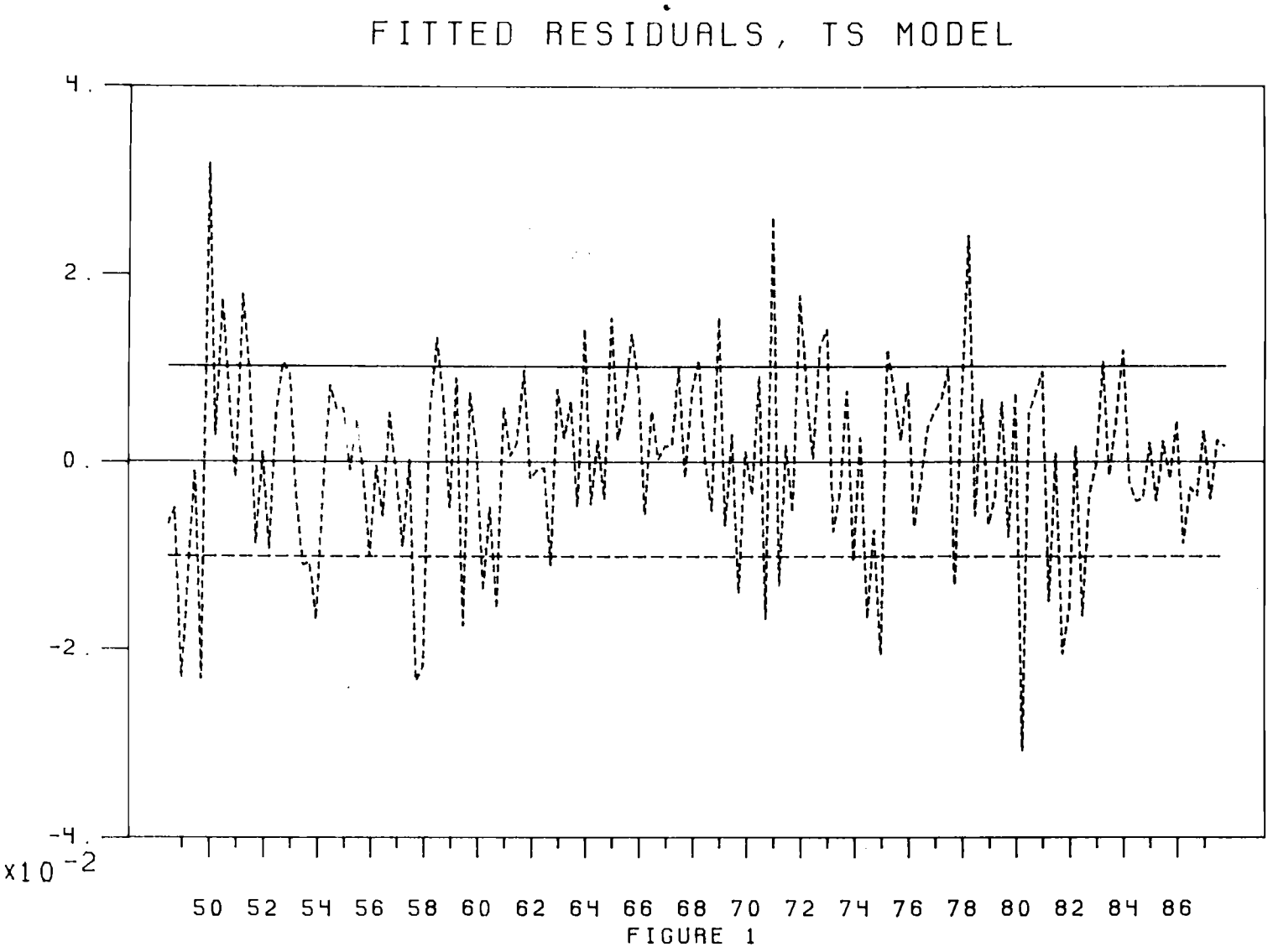
Nelson, Charles R., and Charles I. Plosser,1982, "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," Journal of Monetary Economics, 10, 139–162.

Perron, Pierre, 1987, "The Great Crash, The Oil Price Shock and the Unit Root Hypothesis," Cahier de recherche no 3887, C.R.D.E.,University of Montreal, October.

Prescott, Edward, 1986, "Theory Ahead of Business Cycle Measurement," Federal Reserve Bank of Minneapolis Quarterly Review, Fall.

Rappoport, Peter, and Lucrezia Reichlin, 1987, "Segmented Trends and Nonstationary Time Series," manuscript, Domestic Research Department, Federal Reserve Bank of New York, May.

Sims, Christopher A., 1980, "Macroeconomics and Reality," Econometrica 48, no.1.


FITTED RESIDUALS, TS MODEL
4.
2.
0.
-2.
-4.
x10 -2
50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86
FIGURE 1

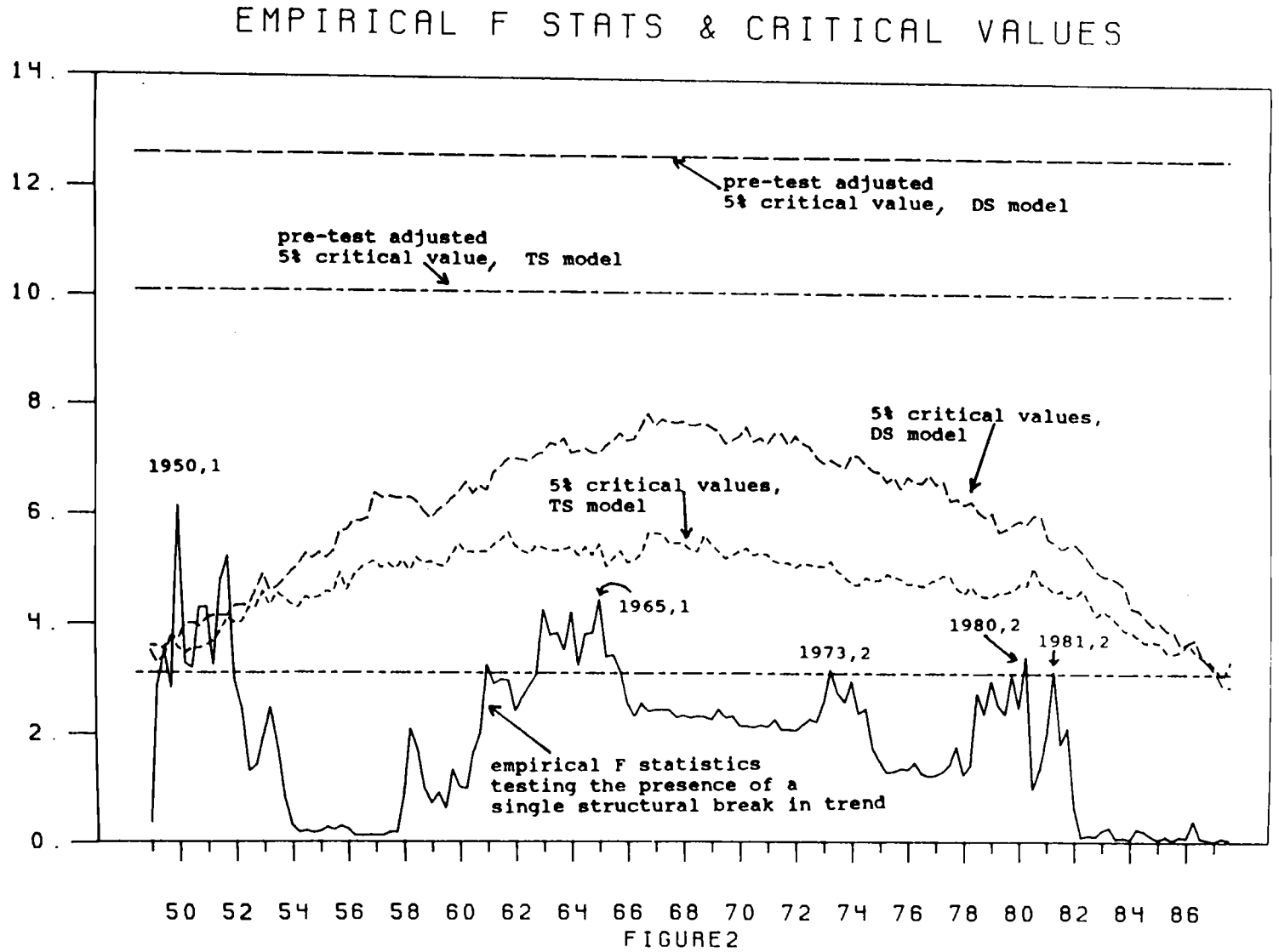
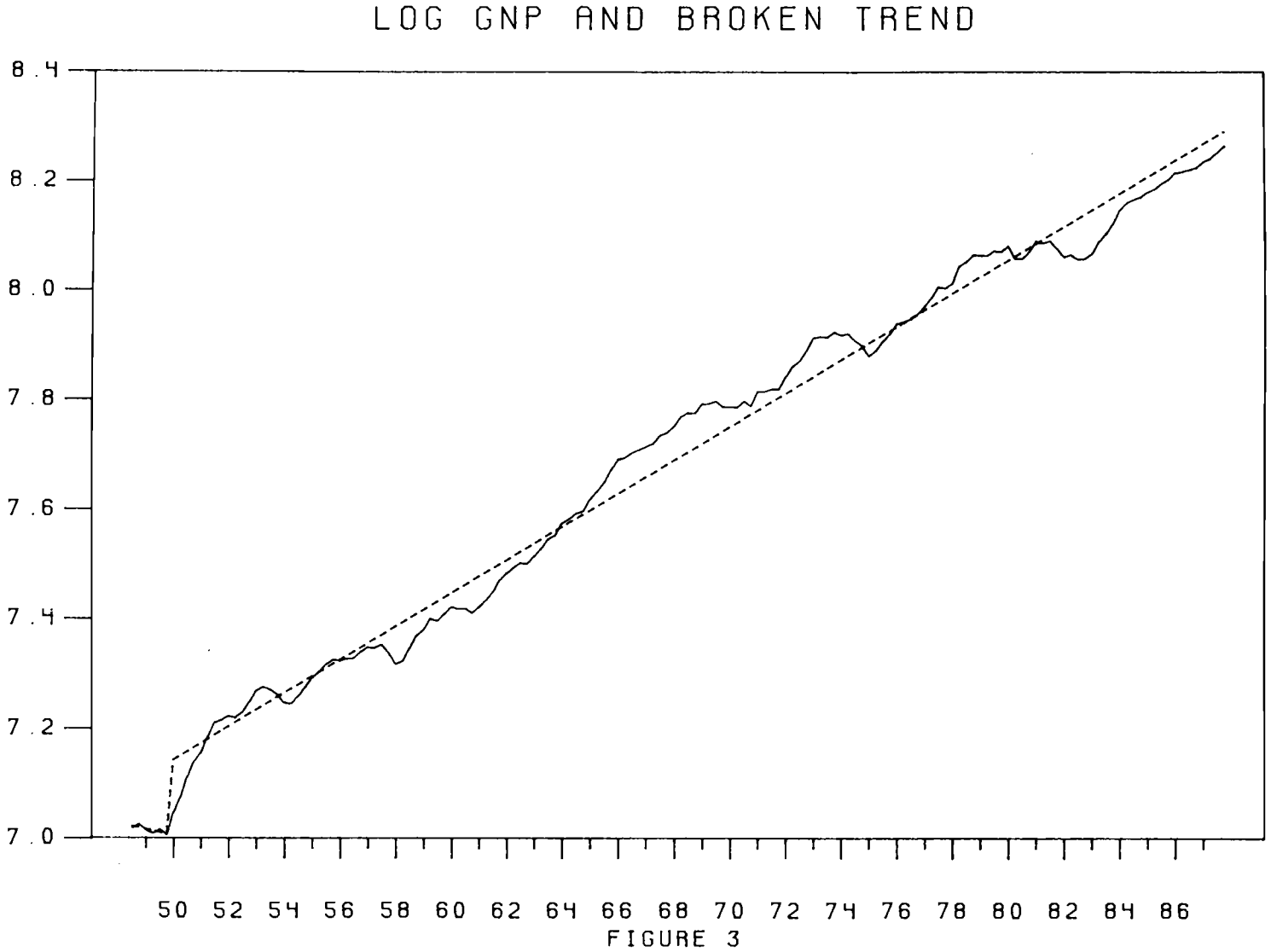

EMPIRICAL F STATS & CRITICAL VALUES
14.
12.
10.
8.
6.
4.
2.
0.
pre-test adjusted
5% critical value, DS model
pre-test adjusted
5% critical value, TS model
5% critical values,
DS model
5% critical values,
TS model
1950,1
1965,1
1973,2
1980,2
1981,2
empirical F statistics
testing the presence of a
single structural break in trend
50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86

FIGURE2


LOG GNP AND BROKEN TREND
8.4
8.2
8.0
7.8
7.6
7.4
7.2
7.0
50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86
FIGURE 3