working paper
department
of economics

ESTIMATION OF STRUCTURAL CHANGE BASED ON
WALD-TYPE STATISTICS

Jushan Bai

94-6 Dec. 1993

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139
ESTIMATION OF STRUCTURAL CHANGE BASED ON WALD-TYPE STATISTICS

Jushan Bai

94-6 Dec. 1993
Estimation of Structural Change Based on Wald-Type Statistics

Jushan Bai*
Massachusetts Institute of Technology†
Revised: December 1993

Abstract

This paper addresses structural change in a linear regression model with an unknown change point. The change point is estimated by maximizing a sequence of Wald-type statistics. This paper focuses on the convergence rate of the estimator. T-consistency is obtained. Asymptotic distributions for the estimated change point and the estimated regression parameters are also considered. The analysis applies to both pure and partial structural changes.

Key words and phrases: Structural change, Wald statistics, weak convergence, change point, T-consistency.
JEL Classification Codes: C13, C21.

*I thank Jerry Hausman and seminar participants at Brown University for very useful comments.
†Mailing Address: E52-274B, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. Phone: (617) 253-6217. Fax: (617) 253-1330. email: jbai@mit.edu.
1 Introduction

In a regression framework, structural change may be considered a regression equation with a shift in some of the regression coefficients. This paper focuses on the inference of the shift point when it is unknown and must be estimated. Estimation of the shift point is based on Wald-type statistics. These statistics are typically used for testing the existence of a structural change; see Chow (1960) for known shift point and Andrews (1993) for unknown shift point. When the shift point is not known, a sequence of Wald statistics is used in constructing the test. This sequence results from sequential estimation carried out for every possible sample split. The shift-point estimator is then defined as the point where the Wald statistic achieves its global maximum. For ease of reference, we shall call this estimator the W-estimator. Under normality with independent and identically distributed errors, the Wald statistic is just the F-statistic. When a single parameter changes, the W-estimator is identical to maximizing a sequence of t-statistics in absolute value.

The Wald type statistics can be thought of as a measure of the distance between the estimator of pre-shift parameters and that of post-shift parameters. Correctly identifying the shift point will tend to maximize this distance and vice versa. One may also consider the pre-shift sample drawn from one distribution and post-shift from another whereby the Wald statistics can be viewed as "between sample variance." Correctly classifying the observations into two samples will maximize the "between samples variance" relative to "within samples variance," as in the classification framework. It is interesting to note and is straightforward to prove that, when the Wald-statistics are computed using a sequence of least squares estimators, the W-estimator is equivalent to minimizing the sum of squared residuals. Furthermore, because LM-type statistics and LR statistics are monotonic transformations of Wald statistics in linear models, the W-estimator can also be obtained by maximizing a sequence of these latter statistics. This may not be true for nonlinear models or for estimation methods other than least squares.
There are at least two advantages for basing the estimation on Wald-type statistics. First and most important, Wald statistics are particularly convenient for the consistency proof given later. We explore the structure of the Wald statistic as a distance between pre-shift and post-shift estimated parameters and are able to show that this distance is globally maximized only near the true shift point. The second advantage is the computational convenience. Wald statistics are easy to compute and are built into many software packages. In addition, testing for a parameter shift and estimating the shift point can be performed concurrently. Upon rejecting the null hypothesis of no change, one may wish to estimate a shift point simply from the test statistics. Testing for a change using Wald statistics is studied by Hawkins (1987) and Andrews (1993).

Structural change in linear regressions was considered early on by Quandt (1958). The problem has received considerable attention recently in the literature. Various tests statistics have been put forward in use. Well known examples include the optimal test of Andrews and Ploberger (1992) and the fluctuation test of Ploberger, Kramer and Kontrus (1989), in addition to the Wald type of tests mentioned earlier. Tests for change in nonstationary time series models have been developed by Perron (1989), Banerjee, Lumsdaine and Stock (1992), Chu and White (1992), Hansen (1992), and Vogelsang (1992), among others. Hackl and Westlund (1989) offer a comprehensive bibliography on the subject of its early development. The estimation of structural change has also been examined by many researchers. A survey is given by Krishnaiah and Miao (1988). Methods of estimation include maximum likelihood (MLE), nonparametric, least squares (LS), least absolute deviation (LAD), and Bayesian estimation, among others. MLE is studied the most, examples including Hinkley (1970), Picard (1985) and Yao (1988). These authors deal with i.i.d, or autoregressive models with a shift. Picard (1985) and Yao (1987) focus on local shifts. Nonparametric estimation is considered by Duembgen (1991) for two i.i.d. samples. LS is studied by Hawkins (1986) and Bai (1993a) and LAD by
Bai (1993b). Bayesian estimation is summarized in the monograph by Broemeling and Tsurumi (1987) (also see Zivot and Phillips (1992)). Further references can be found in Bai (1992). In this paper, we estimate the shift point by using Wald type statistics. These statistics are based on least squares estimation. Our aim is to establish T-consistency for the proposed estimator. Despite the large body of literature, T-consistency has not been obtained for linear regressions.

Furthermore, we consider the problem in the context of partial structural change in which some of the regression parameters hold constant throughout the sample. Thus these parameters should be estimated using the entire sample in order to gain efficiency. The difficulty of the consistency proof increases dramatically under partial structural change in contrast to pure structural change. Bai (1993a) essentially considers a pure change problem and offers a simple argument of consistency. Despite the increase in difficulty, we shall work with a partial structural change model because this model includes pure structural change as a special case. We deal with the pure and partial problems in a unified way by concentrating out the unchanged parameters. The Wald type of statistics serves this purpose well and allows us to focus on the shifted parameters. Least squares method is used to estimate the pre-shift and post-shift parameters and Wald statistics are based on these estimators. The use of least squares estimation enables us to relax the assumption of a known error distribution function as is necessary for maximum likelihood estimation. We also allow heterogeneous and dependent disturbances. The design of the regressors is virtually arbitrary but satisfies some standard assumptions under least squares estimation. In particular, stochastic regressors and time trends are allowed. In addition to the T-consistency of the shift point estimator, root-T consistency for the estimated pre-shift and post-shift parameters is also established.

Asymptotic distribution for the W-estimator is also considered in this paper. This problem is studied for shifts with shrinking magnitudes. Under a shift with a fixed magnitude, it is mathematically intractable to obtain the asymptotic distribution
except for some i.i.d. samples with shift, as shown by Hinkley (1970). The asymptotic distribution allows one to construct confidence intervals for the shift point. We also derive the asymptotic distribution for the estimated regression parameters. We show that the asymptotic distribution for the latter is normal and is the same as if the shift point were known.

This paper is organized as follows. Section 2 specifies the model and the underlying assumptions. Section 3 proves the $T$-consistency of the $W$-estimator. The root-$T$ consistency and asymptotic normality for the regression parameters are also derived in this section. Section 4 examines the asymptotic distribution of the shift-point estimator. Concluding remarks are provided in Section 5. Some technical matters are collected in the Appendix.

2 Models and Assumptions

Consider the following linear regression with a structural change:

$$y_t = x'_t \beta + \epsilon_t, \quad (t = 1, 2, \ldots, k_0) \tag{1}$$
$$y_t = x'_t \beta + z'_t \delta + \epsilon_t, \quad (t = k_0 + 1, \ldots, T) \tag{2}$$

where $\{\epsilon_t\}$ is a sequence of unobservable disturbances that are weakly dependent; $x_t$ and $z_t$ are $p \times 1$ vectors of regressors ($x_t$ is $p \times 1$ and $z_t$ is $q \times 1$ where $p \leq q$); $k_0$ is the unknown shift point, and $\beta$ and $\delta$ are unknown parameters. When the $\epsilon_t$ are uncorrelated over time, the regressors $x_t$ and $z_t$ may include lagged $y_t$. For dependent disturbances, we shall assume the $\epsilon_t$ are uncorrelated with $x_t$ and $z_t$. When $x_t = z_t$, all parameters of the model shift at time $k_0$. When $z_t$ is a sub-vector of $x_t$, a partial shift model is obtained [see Andrews (1993) for a discussion]. More generally, we assume $z_t = R'x_t$ for some matrix $R$ with full column rank so that $z_t$ is a linear transformation of $x_t$. The purpose is to estimate $\beta$, $\delta$, $k_0$, and $\sigma^2$ with a focus on $k_0$.

Let us introduce some further notation. Let $y = (y_1, \ldots, y_T)'$, $X = (x_1, x_2, \ldots, x_T)'$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)'$, $X_1 = (x_1, x_2, \ldots, x_k, 0, \ldots, 0)'$, $X_2 = (0, \ldots, 0, x_{k+1}, \ldots, x_T)'$, and $X_0 = (0, \ldots, 0, x_{k_0+1}, \ldots, x_T)'$. The matrices $X_1$ and $X_2$ depend on $k$, but this
dependence will not be displayed for notational succinctness. Furthermore, let \( Z_1 = X_1 R, \ Z_2 = X_2 R, \) and \( Z_0 = X_0 R. \) Equations (1) and (2) can then be rewritten as
\[
y = X\beta + Z_0\delta + \epsilon. \tag{3}
\]

We consider the problem of testing \( \delta = 0 \) based on the Wald test. Because \( k_0 \) is unknown, the matrix \( Z_0 \) cannot be constructed. The strategy of testing \( \delta = 0 \) is to obtain a sequence of Wald statistics by replacing \( Z_0 \) with \( Z_2 \) for \( k = p, p+1, \ldots, T-p \) (recall \( Z_2 \) depends on \( k \)), where \( p \) is the dimension of \( x_t \), and then using as a test statistic the maximum value of the sequence, or other functionals of this sequence, such as the mean value. For each fixed \( k \), let \( \hat{\beta}_k \) and \( \hat{\delta}_k \) be the least squares estimators of \( \beta \) and \( \delta \), respectively, obtained by regressing \( y \) on \( X \) and \( Z_2 \). The corresponding Wald statistic is then
\[
W_T(k) = \left(\frac{T-p-q}{q}\right) \frac{\hat{\delta}_k'(Z_2'MZ_2)\hat{\delta}_k}{S(k)}, \quad k = p, p+1, \ldots, T-p, \tag{4}
\]
where \( M = I - X(X'X)^{-1}X \) and \( S(k) \) is the sum of squared residuals. Note that \( \hat{\delta}_k \) and \( S(k) \) can be written as
\[
\hat{\delta}_k = (Z_2'MZ_2)^{-1}Z_2'My
\]
\[
S(k) = \sum_{t=1}^{T} (y_t - x_t\hat{\beta}_k - z_t\hat{\delta}_k I(t > k))^2
\]
where \( I(\cdot) \) is the indicator function. The test statistic for testing \( \delta = 0 \) is
\[
W_T = \sup_{\epsilon T \leq k \leq (1-\epsilon)T} W_T(k) \tag{5}
\]
where \( \epsilon > 0 \) is a small positive number. The limiting distribution of \( W_T \) is nonstandard, see Andrews (1993) for details.

Now assume \( \delta \neq 0 \). Our objective is to estimate the shift point \( k_0 \). We define the shift-point estimator as the location where the maximum of Wald statistic is achieved, namely,
\[
\hat{k} = \arg\max_{p \leq k \leq T-p} W_T(k)
\]
where \( p \) is the number of columns of \( X \). We call \( \hat{k} \) the W-estimator for ease of reference. This approach to estimating a shift has been used in empirical applications. Christiano (1992) looked for potential changes in U.S. GNP using F-statistics. Note the restriction \( k \in [p, T - p] \) is needed to ensure the existence of least squares estimators. However, no restriction of the form \( k \in [cT, (1 - c)T] \) is required.

Although both the numerator and denominator of \( W_T(k) \) of (4) depend on \( k \), it is not difficult to show that \( \hat{k} \) can be obtained by maximizing the numerator alone and can also be obtained by minimizing the denominator alone. We state this simple result as a proposition.

**Proposition 1**

\[
\hat{k} = \arg\max_k W_T(k) = \arg\max_k \delta_k(Z'_2MZ_2)\delta_k = \arg\min_k S(k).
\]

That is, the W-estimator obtained by maximizing Wald-type statistics is the same as minimizing the sum of squared residuals. To see this, let \( \tilde{S} \) denote the sum of squared residuals under restricted estimation \((\delta = 0)\), then the Wald statistic is

\[
W_T(k) = \left( \frac{T - p - q}{q} \right) \left( \frac{\tilde{S} - S(k)}{S(k)} \right).
\]

Because \( \tilde{S} \) does not depend on \( k \) and the Wald statistic is a strictly decreasing transformation of \( S(k) \), it follows immediately that \( \hat{k} = \arg\min_k S(k) \). Or equivalently, \( \hat{k} = \arg\max_k \{ \tilde{S} - S(k) \} \). The proposition then follows from \( \tilde{S} - S(k) = \delta_k(Z'_2MZ_2)\delta_k \) [see Amemiya (1985, p. 31-33)].

For a linear model with least squares estimation, LR and LM statistics are monotonic (increasing) transformations of the Wald statistics, thus W-estimator can also be obtained by maximizing LR and LM statistics. Despite these facts, we shall work with Wald statistics for two reasons. First and foremost, Wald statistics lead naturally to the consistency arguments, as will be seen from the proofs below. Other forms of test statistics do not have this advantage. Second, under estimation methods other than least squares (e.g. instrumental variable estimation or GMM estimation),
the results of Proposition 1 may not hold. In particular, for nonlinear models, the equivalence of Wald, LR and LM statistics no longer holds, but our analysis which is based on Wald statistics has the potential to be extended both to other estimators and nonlinear models as well as simultaneous equations systems. For example, Lo and Newey (1985) and Andrews and Fair (1989) consider structural change problems based on Wald-type test statistics for linear and nonlinear simultaneous equations.

In the case of pure structural change, that is, \( x_t = z_t \), the W-estimator becomes
\[
\hat{k} = \arg\max_k \{((\hat{\beta}_1 - \hat{\beta}_2)'[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-1}(\hat{\beta}_1 - \hat{\beta}_2))
\]
where \( \hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y \) and \( \hat{\beta}_2 = (X_2'X_2)^{-1}X_2'y \). The term inside the braces represents a weighted average of the distance (squared) between the pre-shift and post-shift parameter estimators. The W-estimator maximizes this distance.

In addition to the shift point, we are also interested in the regression parameters. Let \( \hat{\beta} = \hat{\beta}(\hat{k}) \) and \( \hat{\delta} = \hat{\delta}(\hat{k}) \) be the estimators of \( \beta \) and \( \delta \) corresponding to \( \hat{k} \). That is, we replace \( Z_0 \) by \( Z_2 \) with \( k = \hat{k} \) and then estimate model (3). We shall establish subsequently that \( \hat{\beta} \) and \( \hat{\delta} \) are root-\( T \) consistent, asymptotically normal, and have the same limiting distribution as if the shift point \( k_0 \) were known.

In what follows, we use \( o_p(1) \) to denote a sequence of random variables converging to zero in probability and \( O_p(1) \) to denote a sequence which is stochastically bounded. For a sequence of matrices \( B_T \), we write \( B_T = O_p(1) \) if each of its elements is \( O_p(1) \). The notation \( \| \cdot \| \) is used to denote the Euclidean norm, i.e. \( \|x\| = (\sum_{i=1}^p x_i^2)^{1/2} \) for \( x \in \mathcal{R}^p \). For a matrix \( A \), we use \( \|A\| \) to denote the vector-reduced norm, i.e., \( \|A\| = \sup_{x \neq 0} \|Ax\|/\|x\| \).

We make the following assumptions:

A1. \( k_0 = [\tau T] \), where \( \tau \in (0, 1) \) and \([\cdot]\) is the greatest integer function.

A2. The data \( \{(y_{it}, x_{it}, z_{it}); 1 \leq t \leq T, T \geq 1\} \) form a triangular array. For notational simplicity, the subscript \( T \) will be suppressed. In addition, \( z_t = R'x_t \), where \( R \) is \( p \times q \), rank(\( R \)) = \( q \), \( z_t \in \mathcal{R}^q \), \( x_t \in \mathcal{R}^p \), \( q \leq p \).
A3. The matrix $\sum_{i=1}^j x_i x'_i$ is positive definite for large values of $|i - j|$ and $\sup_{j \geq 1} \| \frac{1}{2} \sum_{i=1}^{i+j} x_i x'_i \|$ is stochastically bounded for every fixed $l$. Furthermore, $\| \frac{1}{j} \sum_{i=1}^{i+j} x_i x'_i \|$ is bounded away from zero for large $j$.

A4. $\frac{1}{T} (X'X) \xrightarrow{p} Q_{xx}$, where $Q_{xx}$ is finite and positive definite.

A5. $\frac{1}{\sqrt{T}} \sum_{i=1}^{[T]} x_i \epsilon_t \Rightarrow B(s)$, where $B(s)$ is multivariate Gaussian process with zero mean and covariance matrix

$$E\{B(s_1)B(s_2)\}' = \Omega(s_1 \land s_2)$$

where

$$\Omega(s) = \lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{[T]} \sum_{j=1}^{[T]} E(x_i x'_j \epsilon_i \epsilon_j).$$

The notation "$\Rightarrow$" stands for the weak convergence in $D[0,1]$ with the Skorohod topology, see Pollard (1984) and "$x \land y$" stands for the minimum value of $x$ and $y$.

A6. For some real number $r > 2$ and constant $C > 0$,

$$E \left| \sum_{i=1}^{j} x_i \epsilon_t \right|^r \leq C(j - i)^{r/2} \text{ for all } 1 \leq i \leq j \leq T.$$

Assumption A1 assumes that the shift point is bounded away from the end points. This assumption can be slightly more general, for example, $k_0 = [\tau_T T]$, where $\tau_T \to \tau$ and $\tau_T, \tau \in (0,1)$. Assumption A2 allows for trending regressors written in the form $(t/T)^\ell$ $(\ell \geq 0)$ or, more generally, written as any function of the time trend, $g(t/T)$. Expressing trending regressors in this way avoids a scaling matrix that would otherwise be required when deriving limiting distributions. The first half of A3 requires that a sum involving an increasing number of observations must become positive definite. The latter half is typically implied by the strong law of large numbers. Assumptions A4 and A5 are standard for linear regressions. Note that independence of the errors is not assumed. A5 is satisfied for a wide range of settings, see e.g., Wooldridge and White (1988). Assumption A6 is satisfied for $x_i \epsilon_t$ that are martingales or strongly mixing sequences with some moment conditions, see, e.g., Yokoyama (1980) and Andrews and Pollard (1990).
Define $\hat{\tau} = \hat{k}/T$. We shall show that $\hat{\tau}$ is consistent for $\tau$ and prove that $T(\hat{\tau} - \tau) = O_p(1)$ in the following section.

3 The Consistency of $\hat{\tau}$

In this section, we first establish the consistency of the W-estimator and then obtain its convergence rate. The latter result is also used for obtaining the limiting distribution of the W-estimator as well as the limiting distribution of the estimated regression parameters. First, the consistency result:

**Proposition 2** Under assumptions A1-A5, for any $\epsilon > 0$ and $\eta > 0$, there exists $T_0 > 0$, when $T > T_0$,

$$P(|\hat{\tau} - \tau| > \eta) < \epsilon.$$  

To prove this proposition, we define

$$V_T(k) = \delta'_k(Z'_2MZ_2)\hat{k}.$$  

By Proposition 1, $\hat{k} = \arg\max_k V_T(k)$. To obtain consistency, we examine the global behavior of $V_T(k)$. We shall show that, if $\delta \neq 0$, then with high probability, $V_T(k)$ can only be maximized near $k_0$. The basic idea is to decompose $V_T(k) - V_T(k_0)$ into two parts: a "deterministic" part and a stochastic part. The deterministic part is maximized near $k_0$, whereas the stochastic part is uniformly small (in $k$) relative to the deterministic part. To see this, we first notice that

$$\hat{\delta}_k = (Z'_2MZ_2)^{-1}(Z'_2MZ_0)\delta + (Z'_2MZ_2)^{-1}Z'_2M\epsilon,$$

$$\hat{\delta}_{k_0} = \delta + (Z'_0MZ_0)^{-1}Z'_0M\epsilon,$$

therefore

$$V_T(k) - V_T(k_0) = \delta'_k(Z'_2MZ_2)\hat{k} - \delta'_{k_0}(Z'_0MZ_0)\hat{k}_{k_0}$$

$$= \delta'\{(Z'_0MZ_2)(Z'_2MZ_2)^{-1}(Z'_2MZ_0) - (Z'_0MZ_0)\}\delta$$

$$+ h(k, \delta, \epsilon),$$

(6)  

(7)  

(8)  

9
where
\[ h(k, \delta, \varepsilon) = 2\delta'(Z_0'MZ_2)(Z_2'MZ_2)^{-1}Z_2'M\varepsilon - 2\delta'Z_0'M\varepsilon \]
(9)
\[ + \varepsilon'MZ_2(Z_2'MZ_2)^{-1}Z_2'M\varepsilon - \varepsilon'MZ_0(Z_0'MZ_0)^{-1}Z_0'M\varepsilon. \]
(10)

Expression (7) constitutes the deterministic part and \( h(k, \delta, \varepsilon) \) constitutes the stochastic part. Denote
\[ X_\Delta = X_2 - X_0 = (0, ..., 0, x_{k+1}, ..., x_{k_0}, 0, ..., 0)' \quad \text{for } k < k_0 \]
\[ X_\Delta = -(X_2 - X_0) = (0, ..., 0, x_{k_0+1}, ..., x_k, 0, ..., 0)' \quad \text{for } k > k_0 \]
and define \( X_\Delta \) to be a zero vector if \( k = k_0 \) so that \( X_2 = X_0 + X_\Delta \text{sign}(k_0 - k) \). When the distinction between the positive and negative signs is immaterial, we simply write \( X_2 = X_0 \pm X_\Delta \). Set \( Z_\Delta = X_\Delta R \), and let
\[ \gamma(k) = \frac{\delta'((Z_0'MZ_0) - (Z_0'MZ_2)(Z_2'MZ_2)^{-1}(Z_2'MZ_0))\delta}{|k_0 - k|}. \]
(11)

When \( k = k_0 \), both the numerator and denominator of \( \gamma(k) \) are zero. In that case we arbitrarily define \( \gamma(k_0) = \delta'\delta \). Note that \( \gamma(k) \) is non-negative because the matrix inside the braces is semi-positive definite. We have the following identity
\[ V_T(k) - V_T(k_0) = -|k_0 - k|\gamma(k) + h(k, \delta, \varepsilon) \quad \text{for all } k. \]
(12)

The shift-point estimator \( \hat{k} \) must satisfy \( V_T(\hat{k}) \geq V_T(k_0) \), or equivalently, \( h(\hat{k}, \delta, \varepsilon) \geq |k_0 - \hat{k}|\gamma(\hat{k}) \). Thus we have
\[ P(\hat{\tau} - \tau > \eta) = P(|\hat{k} - k_0| > T\eta) \]
\[ \leq P\left( \sup_{|k - k_0| > T\eta} |h(k, \delta, \varepsilon)| \geq \inf_{|k_0 - k| > T\eta} |k_0 - k|\gamma(k) \right) \]
\[ \leq P\left( \sup_{T \leq p \leq T - p} |h(k, \delta, \varepsilon)| \geq T\eta \inf_{|k_0 - k| > T\eta} \gamma(k) \right) \]
\[ = P(\gamma_T^{-1} \sup_{T \leq p \leq T - p} T^{-1}|h(k, \delta, \varepsilon)| \geq \eta) \]
where

\[ \gamma_T = \inf_{|k-k_0|>T} \gamma(k) > 0. \]

Lemma A.2 shows that \( \gamma_T \) is positive and bounded away from zero. Thus consistency will follow from

\[ T^{-1} \sup_{p \leq k \leq T-p} |h(k, \delta, \varepsilon)| = o_p(1). \]  

(13)

Next we verify (13). In fact, we shall show that (13) is \( O_p(T^{-1/2} \log(T)) \), a stronger result than needed. For each fixed \( k \), it is not difficult to see that \( h(k, \delta, \varepsilon) \) grows at most at the rate of \( \sqrt{T} \). It follows that \( h \) divided by \( T \) converges to zero in probability. However, this is not enough. We must prove that the maximum value of \( h \) taken over all possible \( k \) grows at a slower rate than \( T \). This is indeed the case. Here is the precise argument. Divided by \( T \), the first term of the RHS of (9) can be written as

\[ 2\delta' \frac{1}{\sqrt{T}} \frac{1}{\sqrt{T}} (Z_0' M Z_2)(Z_2' M Z_2)^{-1/2}(Z_2' M Z_2)^{-1/2} Z_2 M \varepsilon. \]  

(14)

The law of iterated logarithm implies that

\[ \sup_k \| (Z_2' M Z_2)^{-1/2} Z_2 M \varepsilon \| = O_p(\log T). \]  

(15)

where the supremum is taken over all values of \( k \) such that \( p \leq k \leq T \). Next we show \( B_T = \frac{1}{\sqrt{T}} (Z_0' M Z_2)(Z_2' M Z_2)^{-1/2} = O_p(1) \) uniformly in \( k \), or equivalently,

\[ \sup_k B_T B_T^\prime = \sup_k T^{-1} (Z_0' M Z_2)(Z_2' M Z_2)^{-1} (Z_2' M Z_0) = O_p(1). \]

But the above follows from \( (Z_0' M Z_2)(Z_2' M Z_2)^{-1} (Z_2' M Z_0) \leq (Z_0' M Z_0) \) for all \( k \) and \( T^{-1} (Z_0' M Z_0) = O_p(1) \). Thus (14) is \( O_p(T^{-1/2} \log T) \). The second term of (9) divided by \( T \) is \( T^{-1} Z_0' M \varepsilon = O_p(T^{-1/2}) \). From (15), \( \sup_k T^{-1} \varepsilon' M Z_2(Z_2' M Z_2)^{-1} Z_2' M \varepsilon = O_p(T^{-1}(\log T)^2) \), which corresponds to the first term of (10). The last term of (10) divided by \( T \) is obviously \( O_p(T^{-1}) \). Combining these results, we have \( T^{-1} \sup_k |h(k, \delta, \varepsilon)| = O_p(T^{-1/2} \log T) \). We thus establish the consistency of the W-estimator.

It should be pointed out that the consistency is obtained by globally searching the Wald-type statistics. The consistency argument does not require that the searching
be limited within a positive fraction of the data points such that \( k \in [\epsilon T, (1 - \epsilon)T] \).

The implication is that the sequence of Wald type statistics is well behaved even at
the two ends, attaining its global maximum near the true shift point. In a recent
paper by Nunes, Kuan, and Newbold (1993), they prove the consistency but without
obtaining any rate of convergence. Also their arguments require restricted searching.

Next we establish the \( T \)-consistency result:

**Proposition 3** Under assumptions A1-A6, for any \( \epsilon > 0 \), there exists a \( C > 0 \),
such that for large \( T \),

\[
P(|T(\hat{\delta} - \delta)| > C) = P(|\hat{k} - k_0| > C) < \epsilon.
\]

Again because \( V_T(\hat{k}) \geq V_T(k_0) \) by definition, it suffices to show

\[
P\left( \sup_{|k-k_0|>C} V_T(k) \geq V_T(k_0) \right) < \epsilon.
\]

By Proposition 2, for any \( \epsilon > 0 \) and \( a > 0 \), we have \( P(|\hat{k} - k_0| > Ta) < \epsilon \) for large
\( T \). Thus to prove (16), it is sufficient to show that

\[
p_1 = P\left( \sup_{k \in K(C)} V_T(k) \geq V_T(k_0) \right) < \epsilon
\]

where \( K(C) = \{ k : |k - k_0| > C \quad \text{and} \quad Ta \leq k \leq (1 - a)T \} \) for a small number \( a > 0 \).
Finding the maximum value over the set \( K(C) \) amounts to restricted searching,
but this is legitimate only after establishing consistency. Note \( V_T(k) \geq V_T(k_0) \) is
equivalent to

\[
\frac{h(k, \delta, \epsilon)}{|k_0 - k|} \geq \gamma(k).
\]

Thus

\[
p_1 \leq P\left( \sup_{k \in K(C)} \left| \frac{h(k, \delta, \epsilon)}{k_0 - k} \right| \geq \inf_{k \in K(C)} \gamma(k) \right).
\]

By Lemma A.2, \( \inf_{k \in K(C)} \gamma(k) = \gamma_{CT} \), which is bounded away from 0 for large \( C \)
and large \( T \). Thus it suffices to show that for any fixed \( A > 0 \),

\[
p_2 = P\left( \sup_{k \in K(C)} \left| \frac{h(k, \delta, \epsilon)}{k_0 - k} \right| > A \right) < \epsilon
\]

(17)
when $C$ is large. Consider the terms in (10). For $k \in K(C)$, $Z_2$ involves a positive fraction of data (at least $T^a$ observations), thus $(Z_2^t M Z_2 / T)^{-1} = O_p(1)$ by assumptions A2-A4 and $T^{-1/2} Z_2^t M \varepsilon = O_p(1)$ by assumption A5, where $O_p(1)$ is uniform in $k$. Thus $\varepsilon^t M Z_2 (Z_2^t M Z_2)^{-1} Z_2^t M \varepsilon = O_p(1)$ uniformly on $K(C)$. Similarly, $\varepsilon^t M Z_0 (Z_2^t M Z_0)^{-1} Z_2^t M \varepsilon = O_p(1)$. Therefore (10) divided by $|k_0 - k|$ is bounded by $O_p(1)/|k_0 - k| \leq O_p(1)/C$ because $|k_0 - k| \geq C$. Choose $C$ large enough so that $P(O_p(1)/C > A) < \epsilon/3$ for any pre-given $A > 0$.

Next consider (9). Use $Z_2 = Z_0 \pm Z_\Delta$ to deduce that

$$
\delta'(Z_0^t M Z_2) (Z_2^t M Z_2)^{-1} Z_2^t M \varepsilon \\
= \delta' Z_0^t M \varepsilon \pm \delta' (Z_\Delta^t M Z_2) (Z_2^t M Z_2)^{-1} Z_2^t M \varepsilon \\
= \delta' Z_0^t M \varepsilon \pm \delta' Z_\Delta^t M \varepsilon \pm \delta' (Z_\Delta^t M Z_2) (Z_2^t M Z_2)^{-1} Z_2^t M \varepsilon.
$$

(18)

Therefore,

$$
\left| \frac{\delta'(Z_0^t M Z_2) (Z_2^t M Z_2)^{-1} Z_2^t M \varepsilon - \delta' Z_0^t M \varepsilon}{k_0 - k} \right| \\
\leq \left| \frac{\delta' Z_\Delta^t M \varepsilon}{k_0 - k} \right| + \left| \frac{\delta' (Z_\Delta^t M Z_2) (Z_2^t M Z_2)^{-1} Z_2^t M \varepsilon}{k_0 - k} \right| \\
\leq \left| \frac{\delta' Z_\Delta^t M \varepsilon}{k_0 - k} \right| + \left| \frac{\delta' (Z_\Delta^t M Z_2) (Z_2^t M Z_2)^{-1} Z_2^t M \varepsilon}{k_0 - k} \right|. \quad (20)
$$

(21)

Now by assumptions A3 and A4,

$$
\frac{\delta' (Z_\Delta^t M Z_2)}{k_0 - k} = \frac{Z_\Delta^t Z_\Delta}{k_0 - k} - \frac{Z_\Delta^t X_\Delta}{k_0 - k} \left( \frac{X'X}{T} \right)^{-1} \frac{X'Z_2}{T} = O_p(1),
$$

which together with the functional central limit theorem implies that the second term of (21) is $O_p(T^{-1/2})$. Next, $Z_\Delta M \varepsilon = Z_\Delta \varepsilon - Z_\Delta^t X(X'X)^{-1} X' \varepsilon$. Suppose $k < k_0$ (the case of $k \geq k_0$ is similar), then

$$
\frac{1}{k_0 - k} Z_\Delta^t M \varepsilon = \frac{1}{k_0 - k} \sum_{t=k+1}^{k_0} z_t \varepsilon_t - \frac{1}{k_0 - k} \left( \sum_{t=k+1}^{k_0} z_t x_t' \right) (X'X/T)^{-1} (X' \varepsilon / T) \\
= \frac{1}{k_0 - k} \sum_{t=k+1}^{k_0} z_t \varepsilon_t + O_p(T^{-1/2}).
$$
It remains to be shown that for a given $A$, there exists a $C > 0$, such that

$$P \left( \sup_{k < k_0 - C} \left\| \frac{1}{k_0 - k} \sum_{t=k+1}^{k_0} z_t \varepsilon_t \right\| > A \right) < \varepsilon.$$ 

By Lemma A.3, there exists a $L > 0$, such that for any $A > 0$ and $C > 0$,

$$P \left( \sup_{k < k_0 - C} \left\| \frac{1}{k_0 - k} \sum_{t=k+1}^{k_0} z_t \varepsilon_t \right\| > A \right) < \frac{L}{A^{r(Cr/2)}}$$

which is less than $\varepsilon$ when $C$ is large for a given $A$. This proves (17) and therefore Proposition 3.

T-consistency seems to be typical for shift-point estimators, just as root-$T$ consistency is typical for other parameters under regularity conditions. T-consistency is obtained under nonparametric estimation in the case of two i.i.d. samples, see Dumbgen (1991). The same rate of convergence is also established for LAD estimation and least squares estimation in a strongly mixing sequence or a linear process with a shift in mean, see Bai (1993a, b).

Given the convergence rate of $\hat{k}$, it is relatively easy to show that the estimators of $\beta$ and $\delta$ are not only root-$T$ consistent, but asymptotically normal. In estimating the regression parameters, we use $\hat{k}$ in place of $k_0$. Because of the fast rate of convergence, treating $\hat{k}$ as $k_0$ does not affect the limiting distribution of the estimated regression parameters. Recall that $\hat{\beta} = \hat{\beta}(\hat{k})$ and $\hat{\delta} = \hat{\delta}(\hat{k})$. We have

**Corollary 1** Under assumptions A1-A6, together with $\varepsilon_t$ being uncorrelated,

$$\begin{pmatrix} \sqrt{T}(\hat{\beta} - \beta) \\ \sqrt{T}(\hat{\delta} - \delta) \end{pmatrix} \xrightarrow{d} N(0, \sigma^2 V^{-1}),$$

where

$$V = \text{plim} \frac{1}{T} \begin{pmatrix} \sum_{t=1}^{T} z_t z_t' & \sum_{t=k_0}^{T} z_t z_t' \\ \sum_{t=k_0}^{T} z_t z_t' & \sum_{t=k_0}^{T} z_t z_t' \end{pmatrix}.$$  \hspace{1cm} (22)

For serially-correlated disturbances, the variance-covariance matrix of the limiting distribution is given by

$$V^{-1} U V^{-1},$$
where

$$U = \lim_{T} \frac{1}{T} \left( \sum_{i,j \geq 1}^{T} E(z_i x_j' \varepsilon_i \varepsilon_j) - \sum_{i,j \geq k_0}^{T} E(z_i x_j' \varepsilon_i \varepsilon_j) \right) \left( \sum_{i,j \geq k_0}^{T} E(z_i x_j' \varepsilon_i \varepsilon_j) - \sum_{i,j \geq k_0}^{T} E(z_i x_j' \varepsilon_i \varepsilon_j) \right)$$ (23)

Methods developed by Newey and West (1987) and Andrews (1991) can be used to estimate the matrix $U$. When estimating $V$ and $U$, one uses $\hat{k}$ in place of $k_0$. Notice that $\hat{\alpha}$ and $\hat{\beta}$ have the same limiting distributions as if $k_0$ were known. The conclusion here is that, although the estimators of regression coefficients are determined sequentially, confidence intervals for $\alpha$ and $\beta$ can be constructed in the conventional way (based on t-statistics). This conclusion assumes that a shift does in fact exist.

We required that the vector $z_t$ be a sub-vector of $x_t$ or a linear transformation of $x_t$. It may be possible to have a regressor in $z_t$ but not in $x_t$. That is, only after a structural change does a new variable enter into play. The current proof does not allow this scenario. An extension to cover this case requires similar results to Lemma A.1 and Lemma A.2 without assuming $z_t = R' x_t$. One may explore this possibility by using the identity in Lemma A.4. Another solution is to include the new variable in the matrix $M$, which is equivalent to assigning a zero coefficient to the new variable and including it in $x_t$. But this method produces an inefficient estimation of regression parameters and consequently an inefficient estimation of the shift point as well.

4 Asymptotic Distribution and Confidence Sets

We now consider the limiting distribution of the W-estimator for small shifts (shifts of shrinking sizes as $T$ increases). There are two reasons for this. One is the technical intractability for shifts of fixed magnitude. The limiting distribution is highly data dependent for fixed shifts and thus difficult to obtain. The result of Hinkley (1970) indicates that, even for an i.i.d. normal sample with a mean shift, the limiting distribution is enormously complicated. Second, because of this data dependence, the limiting distribution is of less practical importance. The approach we take here is similar to that of Picard (1985) and Yao (1987). We find a limiting distribution for
small changes. This limiting distribution can then be used as an approximation to the underlying distribution for moderate shifts. Of course, the approximation may not be satisfactory for large shifts. Nevertheless, the limiting distribution provides insight on how serial correlation affects the precision of the shift point estimator and in what way the magnitude of the shift and regressors influence the precision of the shift point.

We assume the magnitude of shifts, \(|\delta|\), depends on \(T\) such that

\[
|\delta_T| \to 0, \quad \sqrt{T}|\delta_T| \to \infty. \tag{24}
\]

In the previous section, we treated \(\delta\) as a constant not varying with \(T\) and established:

\[
\hat{k} = k_0 + O_p(1).
\]

With some modifications, we can show

\[
\hat{k} = k_0 + O_p(|\delta_T|^{-2}). \tag{25}
\]

We omit the details here. Similar results may be found in Bai (1993a,b). Given the convergence rate (25), the limiting distribution of \(\hat{k}\) may be obtained by the local weak convergence of \(V_T(k) - V_T(k_0)\) in conjunction with the continuous mapping theorem for the argmax functional. By the local weak convergence we mean the weak convergence when \(k\) varies in a neighborhood of \(k_0\) such that \(k = k_0 + [v\lambda_T^{-2}]\), where \(\lambda_T^2 = O(|\delta_T|^2)\) and \(v\) is a real number in a compact set. Let

\[
K_T(V) = \{k : k = k_0 + [v\lambda_T^{-2}], |v| \leq V\}.
\]

For any given \(V > 0\), we derive the limiting process of

\[
V_T(k_0 + [v\lambda_T^{-2}]) - V_T(k_0)
\]

for \(v \in [-V,V]\). Suppose \(V_T(k_0 + [v\lambda_T^{-2}]) - V_T(k_0) \Rightarrow G(v)\), the continuous mapping theorem then implies \(\lambda_T^2(\hat{k} - k_0) \overset{d}{\to} \text{argmax}_v G(v)\). For a discussion on the continuity of an argmax functional, we refer to Kim and Pollard (1991).

On \(K_T(V)\), we have a more simplified expression for \(V_T(k) - V_T(k_0)\).
Proposition 4 Under assumptions of A1-A6

\[ V_T(k) - V_T(k_0) = -\delta_T'Z_\Delta Z_\Delta \delta_T \pm 2\delta_T Z_\Delta' \varepsilon + o_p(1) \]

where \( o_p(1) \) is uniform on \( K_T(V) \).

The proof is provided in the Appendix. Thus the limiting process of \( V_T(k) - V_T(k_0) \) is determined by \(-\delta_T'Z_\Delta Z_\Delta \delta_T \pm 2\delta_T Z_\Delta' \varepsilon\). We consider two leading cases.

Case 1: Nontrending regressors but possibly correlated errors. Assume that the regressors satisfy

\[
\text{plim}_{T \to \infty} \frac{1}{T} \sum_{i=1}^{[T\varepsilon]} z_i z_i' = sQ \quad \text{and} \quad \text{plim}_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{T} E(z_i z_j' \varepsilon_i \varepsilon_j) = \Omega
\]

for all \( r \). Let \( \lambda_T^2 = \delta_T'Q \delta_T \). Consider the limit of \( \delta_T'Z_\Delta Z_\Delta \delta_T \) for \( k = k_0 + [v \lambda_T^{-2}] \). Now

\[
\delta_T'Z_\Delta' \delta_T = \lambda_T^{-2} \frac{\delta_T'Z_\Delta' \delta_T}{\lambda_T^{-2}}
\]

Since \( Z_\Delta' \delta_T \) involves \( [v \lambda_T^{-2}] \) (the absolute value of \( [v \lambda_T^{-2}] \)) observations of \( z_i \), we have

\[
\frac{Z_\Delta' \delta_T}{\lambda_T^{-2}} \to |v|Q
\]

uniformly in \( v \in [-V, V] \) (note that \( \lambda_T^{-2} \to \infty \)). This implies

\[
\delta_T'Z_\Delta' \delta_T \to |v|
\]

uniformly in \( v \in [-V, V] \).

Next consider the limit of \( \delta_T'Z_\Delta' \varepsilon \). Assume \( v \leq 0 \), that is, \( k = k_0 + [v \lambda_T^{-2}] \leq k_0 \). By the functional central limit theorem of A5 (re-scaled),

\[
\delta_T'Z_\Delta' \varepsilon = \delta_T' \sum_{t=k+1}^{k_0} z_t \varepsilon_t \Rightarrow \phi W_1(-v)
\]

where \( W_1(\cdot) \) is a Brownian motion process on \( [0, \infty) \) with \( W_1(0) = 0 \) and

\[
\phi^2 = \lim_{T \to \infty} \frac{\delta_T'\Omega \delta_T}{\delta_T'Q \delta_T}.
\]
The weak convergence is in the space $D[-V,0]$, the set of cadlag functions defined on $[-V,0]$, endowed with the Skorohod topology [see Pollard (1984)]. Similarly, when $k \geq k_0$ (i.e., $v \geq 0$), we have

$$\delta_T Z^\Delta \epsilon = \delta_T \sum_{t=k_0+1}^k z_t \epsilon_t \Rightarrow \phi W_2(v)$$

where $W_2(v)$ is another Brownian motion process on $[0, \infty)$ with $W_2(0) = 0$. The two processes $W_1$ and $W_2$ are independent because they are the limits of non-overlapping disturbances that are only weakly dependent. Let $\hat{W}(v) = W_1(-v)$ for $v < 0$, $W(v) = 0$, and $W(v) = W_2(v)$ for $v > 0$, a two-sided Brownian motion process on $(-\infty, +\infty)$. Then combining these results, we have

$$V_T(k_0 + [v\lambda_T^2]) - V_T(k_0) \Rightarrow -|v| + 2\phi W(v).$$

From the continuous mapping theorem for the argmax functional,

$$\lambda_T^2 (\hat{k} - k_0) \xrightarrow{d} \text{argmax}_v \{-|v| + 2\phi W(v)\} = \phi^2 \text{argmax}_v \{W(v) - |v|/2\}.$$ 

The last equality follows from a change in variable. Denote the random variable $\text{argmax}_v \{W(v) - |v|/2\}$ by $V^*$, its density function is given in (35). In view of the definition for $\lambda_T^2$ and $\phi^2$, we can write

$$(\delta^*_T Q \delta_T)^2 (\delta^*_T \Omega \delta_T)^{-1} (\hat{k} - k_0) \xrightarrow{d} V^*$$

(26)

For uncorrelated errors, because $\Omega = \sigma^2 Q$, (26) becomes

$$\frac{(\delta^*_T Q \delta_T)}{\sigma^2} (\hat{k} - k_0) \xrightarrow{d} V^*$$

(27)

A special case is a shift in the intercept only. In this situation, $z_t = 1$, so $Q = 1$. It follows that $\lambda_T^2 (\hat{k} - k_0) \xrightarrow{d} \sigma^2 V^*$.

**Case 2: Trending regressors.** Assume $z_t = g(t/T) = (g_1(t/T), ..., g_q(t/T))'$ and the functions $g_i(x)$ have bounded derivatives on $[0, 1]$. Let $\lambda_T^2 = \delta_T g(\tau)g(\tau)'\delta_T$ which is $O(||\delta_T||^2)$. Let us first show that

$$\delta_T Z^\Delta Z^\Delta \delta_T \rightarrow |v|$$

18
uniformly in \( v \in [-V, V] \). Consider the case of \( v \leq 0 \) (i.e. \( k \leq k_0 \)). Then

\[
\delta'_T Z'_\Delta Z_\Delta \delta_T = \delta_T \sum_{t=k+1}^{k_0} g(t/T)g(t/T)'\delta_T
\]

\[
= (k_0 - k)\delta_T g(\tau)g(\tau)'\delta_T
\]

\[
+ \delta'_T \sum_{t=k+1}^{k_0} [g(t/T) - g(\tau)][g(t/T) - g(\tau)]'\delta_T
\]

\[
+ 2\delta'_T \sum_{t=k+1}^{k_0} [g(t/T) - g(\tau)]g(\tau)'\delta_T
\]

Expression (29) is equal to \((k_0 - k)\lambda_T^2 = -[v \lambda_T^{-2}] \lambda_T^2\), which converges to \(-v\) uniformly in \( v \in [-V, 0] \). We next argue that (30) and (31) are \( o_p(1) \) uniformly on \( K_T(V) \).

Note that (30) is bounded by

\[
\sup_x \left\| \frac{dg(x)}{dx} \right\|^2 \delta_T \delta_T \sum_{t=k+1}^{k_0} (t - k_0)^2/T^2 \leq B_1 \|\delta_T\|^2 (k_0 - k)^3/T^2 \leq B_2 (T^2 \|\delta_T\|^4)^{-1} \rightarrow 0,
\]

for some constants \( B_1 \) and \( B_2 \). The proof that (31) is also \( o_p(1) \) is similar.

Next

\[
\delta_T Z'_\Delta \varepsilon = \delta_T \sum_{t=k+1}^{k_0} g(t/T)\varepsilon_t = \delta_T g(k_0/T) \sum_{t=k+1}^{k_0} \varepsilon_t + \delta'_T \sum_{t=k+1}^{k_0} [g(t/T) - g(k_0/T)]\varepsilon_t.
\]

Suppose the \( \varepsilon_t \) are uncorrelated, then the variance of the second term on the right hand side is

\[
\sigma^2 \delta'_T \sum_{t=k+1}^{k_0} [g(t/T) - g(k_0/T)][g(t/T) - g(k_0/T)]'\delta_T
\]

which converges to zero by (30) and (32). Thus the limiting distribution of (33) is determined by \( \delta'_T g(\tau) \sum_{t=k_1+1}^{k_0} \varepsilon_t \). Because \( k = k_0 + [v \lambda_T^{-2}] \) and \( \lambda_T^2 = \delta_T g(\tau)g(\tau)'\delta_T \), by the functional central limit theorem (see A5),

\[
\delta_T g(\tau) \sum_{t=k_1+1}^{k_0} \varepsilon_t \Rightarrow \sigma W_1(-v).
\]

The argument for \( v \geq 0 \) is similar. In particular,

\[
\delta'_T Z'_\Delta \varepsilon \Rightarrow \sigma W_2(v)
\]
where \( W_2(\cdot) \) is another Brownian motion process independent of \( W_1(\cdot) \). Define \( W(v) \) as in the previous case, we have
\[
V_T(k_0 + [v \lambda_T^2]) - V_T(k_0) \Rightarrow -|v| + 2\sigma W(v).
\]
This implies, by the continuous mapping theorem and a change in variable,
\[
\frac{\lambda_T^2 (\hat{k} - k_0)}{\sigma^2} \xrightarrow{d} \text{argmax}_v \{W(v) - |v|/2\}, \tag{34}
\]
where \( \lambda_T^2 = \delta_T g(\tau)g(\tau)' \delta_T \). For stationary and serially correlated errors, the above convergence still holds but with \( \sigma^2 \) replaced by
\[
\hat{\sigma}^2 = \lim_{h \to \infty} \frac{1}{h} \sum_{i=1}^{h} \sum_{j=1}^{h} E(\varepsilon_i \varepsilon_j).
\]

Confidence Intervals. To construct confidence intervals for \( k_0 \), we use the asymptotic results (26), (27), and (34). Under case 1 (non-trending regressors) together with uncorrelated errors, the limiting distribution of \( \hat{k} \) is given by (27). The distribution function for the random variable \( V^* \xrightarrow{d} \text{argmax}_v \{W(v) - |v|/2\} \) is,
\[
G(x) = 1 + (2\pi)^{-1/2} \sqrt{2e^{-x^2/2}} - \frac{1}{2} (x + 5) \Phi(-\sqrt{x}/2) + \frac{3}{2} e^x \Phi(-3\sqrt{x}/2) \tag{35}
\]
for \( x > 0 \) and \( G(x) = 1 - G(-x) \) [see Yao (1987)], where \( \Phi(x) \) is the distribution function of a standard normal random variable. So quantiles for this random variable are easy to obtain. All we need are estimates for \( \lambda_T^2 = \delta_T Q \delta_T \) and for the variance \( \sigma^2 \). However, \( \delta_T \) is estimable, see Corollary 1. The matrix \( Q \) can be simply taken as \( \frac{1}{T} \sum_{t=1}^{T} z_t z_t' \). The variance \( \sigma^2 \) can be estimated by \( \hat{\sigma}^2 = \frac{1}{T} S(\hat{k}) \). Similar to the result of Corollary 1, it is not difficult to show \( \hat{\sigma}^2 \) is a root-\( T \)-consistent estimator of \( \sigma^2 \). For serially correlated errors, one also needs to estimate the matrix \( \Omega \) and the methods of Newey and West (1987) and Andrews (1991) can be used. For trending regressors, \( \lambda_T^2 = \delta_T g(\tau)g(\tau)' \delta_T \). The matrix \( g(\tau)g(\tau)' \) can be estimated by \( g(\hat{k}/T)g(\hat{k}/T)' \), which only uses the \( \hat{k} \)-th observation. All quantities necessary for constructing confidence intervals are easy to obtain. A 100(1 - \( \alpha \))% confidence interval is given by
\[
[\hat{k} - \hat{\sigma}^2 c_{\alpha/2} / \hat{\lambda}_T^2, \hat{k} + \hat{\sigma}^2 c_{\alpha/2} / \hat{\lambda}_T^2]
\]
20
where $c_{\alpha/2}$ is the $\alpha/2$ percentile of the distribution of $V^*$.

Because the estimated change point $\hat{k}$ is close to $k_0$, one can estimate $\lambda_T^2$ in either case by

$$\hat{e}_T \left( \frac{1}{2m} \sum_{t=k-m}^{k+m} z_t' z_t \right) \hat{e}_T$$

for some $m > 0$. In the case of trending regressor such that $z_t = g(t/T)$, it is not difficult to show that $\frac{1}{2m} \left( \sum_{t=k-m}^{k+m} z_t' z_t \right) \rightarrow g(\tau)g(\tau)'$, for each fixed $m$. This is also true for $m$ growing unbounded but satisfying $m/T \rightarrow 0$; for example, $m = \sqrt{T}$. If $z_t$ contains no trending regressors, then $\frac{1}{2m} \left( \sum_{t=k-m}^{k+m} z_t' z_t \right)$ should be close to $Q$ provided $m$ is relatively large. Finally, the confidence intervals for regression parameters $\beta$ and $\delta$ are constructed in the usual way.

5 Concluding Remarks

We have considered the structural change problem in a linear regression model where part or all of the coefficients have a shift occurring at an unknown time. The unknown shift point is estimated by maximizing a sequence of Wald statistics. We established the $T$-consistency for the estimated shift point. In the case of partial structural change, the unchanged regression parameters are estimated with the whole sample, whereas the shifted parameters are estimated with subsamples. With the presence of structural change, we have demonstrated that the sequence of Wald type statistics are well behaved in the sense that the sequence achieves its global maximum near the true shift point. Thus the maximization is done via global searching. Confining the search for a maximum in the middle of the sequence (ignoring some fractions of the sequence in the beginning or at the end) is not necessary. We also noted that for a linear regression with least squares estimation the shift-point estimator can be equivalently obtained by maximizing a sequence of F-statistics, LM-statistics, or LR-statistics. When only a single parameter is allowed to change, the estimator can be obtained by maximizing a sequence of t-statistics in absolute value as well. In
addition, we studied the asymptotic distributions for the shift-point estimator as well as the regression-parameter estimators. These results enable one to construct confidence intervals. Most importantly, the asymptotic results provide insight on how the precision of the shift-point estimator depends on the regressors, the magnitude of the shift, and serial correlations in the data. Our results hold for a wide range of regressor designs, with time trend and stochastic regressors as special cases. We also permit heterogeneous and dependent disturbances.

It will be interesting to extend the argument of this paper to the setup considered by Andrews (1993), who employs Wald, LM and LR statistics to test for the existence of a structural change in nonlinear models. These statistics are constructed using instrumental variable estimation or more generally the Hansen-type GMM estimation. By maximizing a sequence of these statistics, one also obtains a shift-point estimator. An important unresolved topic is the consistency of the resulting estimator. Much work is needed to make the analysis of this paper applicable for nonlinear models and for models with nondifferentiable objective functions such as LAD. We hope this paper will stimulate further research in this area.
A Appendix

Lemma A.1 The following two inequalities hold:

\[
(Z_0'MZ_0) - (Z_0'MZ_2)(Z_2'MZ_2)^{-1}(Z_2'MZ_0) \\
\geq R'(X_\Delta^X X_\Delta)(X_2'X_2)^{-1}(X_0'X_0)R, \quad k < k_0 \tag{36}
\]

\[
(Z_0'MZ_0) - (Z_0'MZ_2)(Z_2'MZ_2)^{-1}(Z_2'MZ_0) \\
\geq R'X_\Delta^X X_\Delta(X'X - X'_X X_0)^{-1}(X'X - X'_X X_0)R, \quad k \geq k_0. \tag{37}
\]

Proof of (36). Write \( H = (X_2'X_2)^{-1} - (X'X)^{-1} \). For \( k \leq k_0 \),

\[
(Z_0'MZ_2)(Z_2'MZ_2)^{-1}(Z_2'MZ_0) = \tag{38}
R'(X_0'X_0)H(X_2'X_2) \left\{ R(X_2'X_2)H(X_2'X_2)R \right\}^{-1} R'(X_2'X_2)H(X_0'X_0)R. \tag{39}
\]

Let \( A = H^{1/2}(X_2'X_2)R \). Since \( I - A(A'A)^{-1}A' \) is a projection matrix, we have \( I - A(A'A)^{-1}A' \geq 0 \). Multiplying this inequality by \( R'(X_0'X_0)H^{1/2} \) from the left and multiplying from the right by \( H^{1/2}(X_0'X_0)R \), we obtain

\[
R'(X_0'X_0)H(X_0'X_0)R - R'(X_0'X_0)H^{1/2}A(A'A)^{-1}A'H^{1/2}(X_0'X_0)R \geq 0.
\]

The second term above is identical to (39). Thus it suffices to show

\[
Z_0'MZ_0 - R'(X_0'X_0)H(X_0'X_0)R \geq R'(X_\Delta^X X_\Delta)(X_2'X_2)^{-1}(X_0'X_0)R. \tag{40}
\]

In fact, the equality holds in (40) because the left hand side of (40) is

\[
R'(X_0'X_0) \left[ (X_0'X_0)^{-1} - (X'X)^{-1} - H \right] (X_0'X_0)R \\
= R'(X_0'X_0)[(X_0'X_0)^{-1} - (X_2'X_2)^{-1}](X_0'X_0)R \\
= R'(X_\Delta^X X_\Delta)(X_2'X_2)^{-1}(X_0'X_0)R
\]

The last equality follows from \( X_2'X_2 = X_0'X_0 + X_\Delta^X X_\Delta \). The proof of (37) is similar and is omitted.
Lemma A.2 Under assumptions A2-A4, for large $C > 0$, $\gamma_{CT}$ is positive and bounded away from zero, where

$$\gamma_{CT} = \inf_{|k-k_0| > C} \gamma(k)$$

and $\gamma(k)$ is defined in (11). That is, $\liminf_{T \to \infty} \gamma_{CT} = \gamma > 0$ in probability. Also notice that for any number $\eta > 0$,

$$\gamma_T = \inf_{|k-k_0| > T\eta} \gamma(k) = \inf_{|k-k_0| > C} \gamma(k)$$

because $T\eta > C$ for large $T$.

Proof: For $k \leq k_0$, by Lemma A.1 part (i),

$$\gamma(k) \geq \frac{\delta'R'X_{\Delta}X_{\Delta}}{k_0 - k}(X_2'X_2)^{-1}(X_0'X_0)R\delta.$$  \hspace{1cm} (41)

If $X_{\Delta}X_{\Delta}$ is positive definite then the RHS of (36) is positive definite since it can be written as $R'[X_0'X_0]^{-1} + (X_{\Delta}X_{\Delta})^{-1}]^{-1}R$. Thus if $X_{\Delta}X_{\Delta} > 0$, the LHS of (41) is positive. Note that $X_{\Delta}X_{\Delta}$ is in fact positive definite, when $k$ and $k_0$ are far apart (i.e., when $X_{\Delta}$ contains enough nonzero elements). Also when $k \leq k_0,

\frac{1}{k_0 - k}X_{\Delta}X_{\Delta} = \frac{1}{k_0 - k} \sum_{i=k+1}^{k_0} x'_ix_i$ is bounded away from zero, by A2 and A3, for all $k$ such that $k_0 - k$ is large. In addition,

$$(X_2'X_2)^{-1}(X_0'X_0) = \frac{T-k_0}{T-k} \left( \frac{X_2'X_2}{T-k} \right)^{-1} \frac{X_0'X_0}{T-k}$$

is bounded away from zero for all $k \leq k_0$. Thus we can choose $C$ sufficiently large so that

$$\inf_{|k-k_0| > C} \left( \frac{\delta'R'X_{\Delta}X_{\Delta}}{k_0 - k}(X_2'X_2)^{-1}(X_0'X_0)R\delta \right)$$

is bounded away from zero. The case for $k \geq k_0$ can be proved similarly.

Lemma A.3 Under assumption A6, there exists $L > 0$ such that for any $A > 0$,

$$P \left( \sup_{j \geq m} \left| \sum_{t=1}^{j} z_t \varepsilon_t \right| > A \right) \leq \frac{L}{m^{r/2}A^r}$$
Proof: This is essentially proved by Serfling (1970) (Theorem 5.1.). Details are omitted. Note for $z_t \epsilon_t$ i.i.d. or martingale differences, the lemma is the Hajek and Renyi (1955) inequality. Extension of the Hajek and Renyi inequality to a linear process of martingale differences is given in Bai (1993a).

**Proof of Corollary 1.** Let us use $\hat{Z}_0$ to denote $Z_0$ when $k_0$ is replaced by $\hat{k}$. Then the estimators $\hat{\beta}$ and $\hat{\delta}$ are obtained by regressing $y$ on $X$ and $\hat{Z}_0$. Equation (3) can be written as

$$y = X\beta + \hat{Z}_0\delta + \epsilon^*$$

with $\epsilon^* = \epsilon + (Z_0 - \hat{Z}_0)\delta$. Proceeding in the usual way, we have

$$\sqrt{T} \left( \begin{array}{c} \hat{\beta} - \beta \\ \hat{\delta} - \delta \end{array} \right) = \left[ \frac{1}{T} \left( \begin{array}{cc} X'X & X'\hat{Z}_0 \\ \hat{Z}_0X & \hat{Z}_0\hat{Z}_0 \end{array} \right) \right]^{-1} \frac{1}{\sqrt{T}} \left( \begin{array}{c} X'\epsilon + X'(Z_0 - \hat{Z}_0)\delta \\ \hat{Z}_0\epsilon + \hat{Z}_0(Z_0 - \hat{Z}_0)\delta \end{array} \right).$$

All we have to show is that the limit of the right-hand side is the same as the limit when $\hat{Z}_0 = Z_0$. Let us show

$$\text{plim} \frac{1}{\sqrt{T}} X'(Z_0 - \hat{Z}_0)\delta = 0. \quad (42)$$

Supposing $\hat{k} \leq k_0$ (the case of $\hat{k} > k_0$ is similar), we have

$$\frac{1}{\sqrt{T}} X'(Z_0 - \hat{Z}_0) = \frac{1}{\sqrt{T}} \sum_{t=\hat{k}+1}^{k_0} x_t z_t'. \quad (43)$$

Since the sum only involves $k_0 - \hat{k}$ terms, and $k_0 - \hat{k} = O_p(1)$ by Proposition 3, (43) converges to zero. The zero limit for (43) certainly implies $\text{plim} \frac{1}{T} X'\hat{Z}_0 = \text{plim} \frac{1}{T} X'Z_0$. All other terms involving $\hat{Z}_0$ can be treated similarly and the derivations are omitted.

**Remarks:** Corollary 1 not only holds for fixed $\delta$, but also holds when $\|\delta_T\| \to 0$ and $\sqrt{T}\|\delta_T\| \to \infty$, as in the setup of Section 4. In this case, $\hat{k} = k_0 + \|\delta_T\|^{-2} O_p(1)$, and (42) can be proved as follows. Note:

$$\frac{1}{\sqrt{T}} X'(Z_0 - \hat{Z}_0)\delta_T \leq \frac{1}{\sqrt{T}\|\delta_T\|} \sum_{t=\hat{k}+1}^{k_0} x_t z_t' \|\delta_T\|^2.$$

Since the sum involves about $\|\delta_T\|^{-2}$ terms, $\|\sum_{t=\hat{k}+1}^{k_0} x_t z_t'\| \|\delta_T\|^2 = O_p(1)$. By assumption, $(\sqrt{T}\|\delta_T\|)^{-1} \to 0$, and so the desired result follows.
Lemma A.4 The following identity holds

\[ \delta_T' \{ (Z_0' M Z_0) - (Z_0' M Z_2)(Z_2' M Z_2)^{-1} (Z_2' M Z_0) \} \delta_T = \]

\[ \delta_T' (Z_\Delta' M Z_\Delta) \delta_T - \delta_T (Z_\Delta' M Z_2)(Z_2' M Z_2)^{-1} (Z_2' M Z_\Delta) \delta_T. \]

The proof follows simplify from the fact that \( Z_0' M Z_2 = Z_2' M Z_2 \pm Z_\Delta' M Z_2 \).

Lemma A.5 Under the assumptions of A1-A6,

(i) \( \| T^{-1/2} \delta_T' (Z_\Delta' X) \| = o_p(1) \)

(ii) \( \| T^{-1/2} \delta_T' (Z_\Delta' M Z_2) \| = o_p(1) \)

(iii) \( \| T^{-1/2} \varepsilon' M Z_\Delta \| = o_p(1) \)

(iv) \( \varepsilon' M Z_2 (Z_2' M Z_2)^{-1} Z_2' M \varepsilon - \varepsilon' M Z_0 (Z_0' M Z_0)^{-1} Z_0' M \varepsilon = o_p(1) \)

where the \( o_p(1) \) is uniform on \( K_T(V) \).

Proof of (i).

\[ \frac{1}{\sqrt{T}} \delta_T' (Z_\Delta' X) = \frac{k_0 - k}{\sqrt{T}} \delta_T' \frac{Z_\Delta' X}{k_0 - k} \]

\[ \leq \frac{D \lambda_T^2}{\sqrt{T}} \delta_T \| O_p(1) = \frac{1}{\sqrt{T} \| \delta_T \|} O_p(1) = o_p(1). \]

Proof of (ii)

\[ \frac{1}{\sqrt{T}} \delta_T' Z_\Delta' M Z_2 = \frac{1}{\sqrt{T}} \delta_T' Z_\Delta' Z_2 - \frac{1}{\sqrt{T}} \delta_T' (Z_\Delta' X) \left( \frac{X'X}{T} \right)^{-1} \frac{X'Z_2}{T}. \]

By part (i), the second term is \( o_p(1) O_p(1) = o_p(1) \). That the first term is \( o_p(1) \) can be proved in exactly the same way as in part (i) (Note that \( Z_\Delta' Z_2 \) equals \( Z_\Delta' Z_\Delta \) for \( k < k_0 \) and equal to zero for \( k \geq k_0 \)).

Proof of (iii).

\[ T^{-1/2} \varepsilon' M Z_\Delta = T^{-1/2} \varepsilon' Z_\Delta - \frac{\varepsilon' X}{\sqrt{T}} \left( \frac{X'X}{T} \right)^{-1} \frac{X'Z_\Delta}{T}. \]
Because $\varepsilon'Z_\Delta$ consists of only $\lambda_T^{-2}$ observations, the functional central limit theorem implies that $\lambda_T \|\varepsilon'Z_\Delta\| = O_p(1)$. This together with $\sqrt{T}\lambda_T \to \infty$ implies that the first term on the right is $o_p(1)$. Consider the second term. Because $X'Z_\Delta/(k_0-k) = O_p(1)$ and $|(k_0-k)/T| = o(1)$, it follows that $(X'Z_\Delta)/T = o_p(1)$, where $O_p(1)$ and $o(1)$ both being uniform on $K_T(V)$. Finally, from $\varepsilon'X/\sqrt{T} = O_p(1)$ and $(X'X/T)^{-1} = O_p(1)$, the second term is $O_p(1) \cdot o_p(1) = o_p(1)$.

Proof of (iv). Use $Z_2 = Z_0 \pm Z_\Delta$ to obtain

$$\varepsilon'MZ_2(Z_2'MZ_2)^{-1}Z_2'M\varepsilon =$$

$$\varepsilon'MZ_0(Z_2'MZ_2)^{-1}Z_0'M\varepsilon + \varepsilon'MZ_\Delta(Z_2'MZ_2)^{-1}Z_2'M\varepsilon + \varepsilon'MZ_0(Z_2'MZ_2)^{-1}Z_\Delta'M\varepsilon.$$

The result of (iii) implies that the last two terms above are $o_p(1)$. Thus the left hand side of (iv) can be written as

$$\varepsilon'MZ_0 \left[ \left( \frac{Z_2'MZ_2}{T} \right)^{-1} - \left( \frac{Z_0'MZ_0}{T} \right)^{-1} \right] \frac{Z_0'M\varepsilon}{\sqrt{T}} + o_p(1).$$

Because the two matrices inside the bracket converge to the same limit on $K_T(V)$, (iv) is proved.

Proof of Proposition 4: By the definition of $V_T(k) - V_T(k_0)$ [see (7) and (8)] and Lemma A.4,

$$V_T(k) - V_T(k_0) = -\delta_T(Z_\Delta'MZ_\Delta)\delta_T + \delta_T(Z_\Delta'MZ_2)(Z_2'MZ_2)^{-1}(Z_2'MZ_\Delta)\delta_T + h(k, \delta_T, \varepsilon). \quad (45)$$

From the definition of $M$, the first term of (45) is

$$\delta_T(Z_\Delta'MZ_\Delta)\delta_T = \delta_T(Z_\Delta'MZ_\Delta)\delta_T - \frac{\delta_TZ_\Delta'X}{\sqrt{T}} \left( \frac{X'X}{T} \right)^{-1} \frac{X'Z_\Delta\delta_T}{\sqrt{T}}.$$

By Lemma (A.5)(i), the second term on the right is $o_p(1)$, leaving $\delta_T(Z_\Delta'MZ_\Delta)\delta_T$. Now the second term of (45) can be written as, upon appropriate scaling,

$$\delta_T(Z_\Delta'MZ_2)(Z_2'MZ_2)^{-1}(Z_2'MZ_\Delta)\delta_T$$

$$= T^{-1/2}\delta_T(Z_\Delta'MZ_2)(Z_2'MZ_2/T)^{-1}(Z_2'MZ_\Delta)\delta_T T^{-1/2}. \quad (46)$$
On $K_T(V)$, $(Z'_2MZ_2/T)^{-1} = O_p(1)$ and by Lemma A.5 (ii), (46) is $o_p(1)$. Next, consider the last term of (45) $h(k, \delta_T, \epsilon)$ [see (9) and (10) for its definition]. Lemma A.5 (iv) says (10) is $o_p(1)$. Using $(Z'_0MZ_2) = (Z'_2MZ_2) \pm Z'_\Delta MZ_2$ together with Lemma A.5, it is easy to show

$$2\delta_T'(Z'_0MZ_2)(Z'_2MZ_2)^{-1}Z'_2M\epsilon - 2\delta_T'Z'_0M\epsilon = \pm 2\delta_T'Z_\Delta\epsilon + o_p(1).$$

Combining these results yields Proposition 4.

**References**


