Statistics and Probability Letters 102 (2015) 22–29

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Time series regression with persistent level shifts

Jonathan Woody

Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762, United States

ARTICLE INFO

Article history:
Received 11 November 2014
Received in revised form 22 March 2015
Accepted 25 March 2015
Available online 3 April 2015

Keywords:
Time series
Changepoints
Linear model
Renewal process

ABSTRACT

A changepoint in a time series is a time in which any change in the distributional form (marginal or joint) of the series occurs. This includes changes in mean or covariance structure of the time series. Mean level shift changepoints have been shown to dramatically influence linear trend estimates obtained from a simple linear regression model. This study provides an asymptotic analysis of a time series regression model experiencing an increasing number of mean level shifts at known times. It is shown that one may consistently estimate any finite number of unknown parameters in a time series polynomial regression, so long as two or more consecutive observations without a changepoint occurs infinity often in the limit.

1. Introduction

This study investigates a time series regression model experiencing persistent level shifts that occur at known times. These level shift times are the changepoints of the model. The statistical advance is that the errors in the model are assumed to be stationary and are not limited to the IID setting. The number of changepoints is allowed to tend to infinity as time increases.

This work is driven by climatological considerations, as climatological series frequently experience mean level shifts which can dramatically influence regression parameter estimates (Lu and Lund, 2007; Solow, 1987). Such shifts may occur whenever a station location, observer, or instrumentation are changed. These times are frequently recorded in the associated station metadata and are considered known in this study. While this problem was investigated with climatological applications in mind, it does fit other applications as well. For example, each time one of the 30 companies is changed in the Dow Jones industrial average, a mean level shift can occur. Since the time of the substitution is reported, it is considered known. Scenarios in quality control also arise where a mean level shift in manufacturing output occurs each time a part or worker is replaced.

While not the focus of this study, the problem of identifying unknown changepoints remains an active area of research. Recent works by Aue and Horváth (2013) and Jandhyala et al. (2013) discuss inference in this setting. In particular, identification of unknown changepoints in polynomial regression designs are discussed in Aue et al. (2008, 2009), Esterby and El-Shaarawi (1981), MacNeill (1978), and Muggeo (2003), for example. While methods for inference in the presence of undocumented changepoints are of interest in the literature (Bai and Perron, 1998; Chen, 2014, are examples), there is still much room for theoretical development of the *documented* changepoint inference problem. This work extends the results in Woody and Lund (2014) to a regression model with a stationary error process.

2. The model

Let $\{Y_t\}$ be a time series sampled from the General Linear Model (GLM)

$$Y_t = \mu + \theta_1 f_1(t) + \cdots + \theta_m f_m(t) + \epsilon_t, \quad t = 1, 2, \ldots, n, \tag{2.1}$$

E-mail address: jwoody@math.msstate.edu.

http://dx.doi.org/10.1016/j.spl.2015.03.011
0167-7152/

where the regression factors $f_1(t), \ldots, f_m(t)$ are known functions of time, $\theta_1, \ldots, \theta_m$ are unknown parameters, μ is an intercept term, and $\{e_t\}$ is a mean zero stationary error process. A vast literature exists on the GLM parameter estimation in (2.1), with a thorough asymptotic treatment in Bickel and Doksum (2001), for example. The model in (2.1) is restrictive in that μ is constant. To accommodate persistent mean level shifts, a version of (2.1) that allows for multiple level shifts is studied:

$$Y_t = \delta_t + \theta_1 f_1(t) + \cdots + \theta_m f_m(t) + \epsilon_t, \quad t = 1, 2, \ldots, n, \tag{2.2}$$

where the $\{\delta_t\}$ term models the mean level shift. For a fixed sample size n and $t \in \{1, 2, \ldots, n\}$

$$\delta_t = \begin{cases} \Delta_1, & 1 \le t < \tau_1 \\ \Delta_2, & \tau_1 \le t < \tau_2 \\ \vdots & \vdots \\ \Delta_{k+1}, & \tau_k \le t < n+1, \end{cases}$$

where $k = k(n)$ denotes the number of changepoints up to time n. Let $\tau_0 < \tau_1 < \cdots < \tau_k$ denote the ordered known level shift times, with $\tau_0 = 1$ and $\tau_{k+1} = n + 1$ by convention. Note that $\tau_0, \ldots, \tau_k$ partition $\{1, \ldots, n\}$ into $k(n) + 1$ regimes. Let r index the rth regime of the data and set $H_r = \{\tau_{r-1}, \ldots, \tau_r - 1\}$ as the set of times when the series is in regime r. Denote the length of the rth regime by

$$\ell_r = \tau_r - \tau_{r-1}, \tag{2.3}$$

for $r = 1, \ldots, k(n) + 1$. If one sets $\delta_{r(t)} = \sum_{r=1}^{k(n)+1} \mathbb{1}_{\{t \in H_r\}} \delta_r$, then (2.2) may be written as

$$Y_t = \delta_{r(t)} + \theta_1 f_1(t) + \cdots + \theta_m f_m(t) + \epsilon_t, \tag{2.4}$$

where $r(t) \in \{1, \ldots, k + 1\}$ is the regime the series experienced at time t. The errors $\{e_t\}_{t=1}^n$ are assumed stationary, with an autocovariance structure specified below. Since $k(n) \to \infty$ as $n \to \infty$, the model accommodates an infinite number of level shifts in the limit. Woody and Lund (2014) construct estimators for $\theta_1, \ldots, \theta_m$ in the model presented in Eq. (2.4) with IID errors as

$$\hat{\mathbf{\Theta}}(n) = [\mathbf{X}^T(n)\mathbf{X}(n)]^{-1}\mathbf{X}^T(n)\mathbf{Y}(n), \tag{2.5}$$

where

$$\mathbf{X}_{t,v}(n) := f_v(t) - \bar{f}_{v,r(t)}, \quad v = 1, 2, \ldots, m;\ t = 1, 2, \ldots, n, \tag{2.6}$$

and

$$\bar{f}_{u,r} = \frac{1}{\ell_r} \sum_{t \in H_r} f_u(t). \tag{2.7}$$

To ease notation, the columns of $\mathbf{X}(n)$ are assumed to be linearly independent for all n. This is tantamount to assuming that the factors in (2.4) are identifiable.

Let $\mathbf{X}_{*,v}(n) = (X_{1,v}, \ldots, X_{n,v})^T$ be an $n \times 1$ vector containing the vth column of $\mathbf{X}(n)$. The notation $\mathbf{X}_v(n) = \mathbf{X}_{*,v}(n)$ is used in what follows. However, the tth element in the ith column will be referenced as $\mathbf{X}_{t,i}(n)$. Define the inner products

$$\langle \mathbf{X}_u(n), \mathbf{X}_v(n) \rangle = \sum_{r=1}^{k+1} \sum_{t \in H_r} (f_u(t) - \bar{f}_{u,r})(f_v(t) - \bar{f}_{v,r}), \quad u, v = 1, 2, \ldots, m. \tag{2.8}$$

For an $n \times 1$ vector $\mathbf{v} = (v_1, \ldots, v_n)^T$, let $\|\cdot\|$ denote the Euclidean squared-norm, formulated as

$$\|\mathbf{v}\| = \left(\sum_{t=1}^n v_t^2 \right)^{1/2}.$$

Additionally, define the Euclidean matrix norm of an $n \times n$ matrix $\mathbf{A}$ by

$$\|\mathbf{A}\| = \max_{\|\mathbf{b}\|=1} \|\mathbf{Ab}\|$$

where b is an $n \times 1$ vector of unit length.

Finally, two more matrices which are introduced in Woody and Lund (2014) will be of use. Let $\mathbf{E}(n)$ denote the diagonal matrix

$$\begin{pmatrix} \langle \mathbf{X}_1(n), \mathbf{X}_1(n) \rangle & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \langle \mathbf{X}_m(n), \mathbf{X}_m(n) \rangle \end{pmatrix}. \tag{2.9}$$