Statistics and Probability Letters 123 (2017) 122–127

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A new variance component score test for testing distributed lag functions with applications in time series analysis

Yin-Hsiu Chen *, Bhramar Mukherjee

Department of Biostatistics, University of Michigan, United States

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 2 December 2016
Accepted 2 December 2016
Available online 16 December 2016

Keywords:
Distributed lag models
Generalized linear mixed model
Likelihood ratio test
Time series
Variance component score test

ABSTRACT

We propose to test a given constrained distributed lag model (DLM) of the form $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\theta}$ against an unconstrained alternative using a variance component score test (VCST) and show that VCST is more powerful than the standard likelihood ratio test in a simulation study.

1. Introduction

Distributed lag models (DLMs), first introduced in the econometrics literature, are often used to model the current value of a response variable at time t in association with both the current value and the lagged values of an independent variable in a time series analysis. For example, environmental epidemiologists model the current day mortality counts in association with daily air pollution related exposure, such as PM_{10}, up to several days prior to the event day (Schwartz, 2000; Welty et al., 2009). Economists study the long-term effects of macroeconomic variables on stock returns using distributed lag models (Majid and Yusof, 2009; Hsu, 2015). Unconstrained DLMs entail the potential problem of multicollinearity among the various lagged values of the independent variable and the number of parameters to be estimated can be large. Constrained DLMs assume some functional relationship between lag coefficients and lag indices (in the form of a distributed lag function) and serve as a potential solution to the problem. Common constraints include a polynomial (Almon, 1965), a spline (Corradi, 1977), and a natural cubic spline (Hastie and Tibshirani, 1993). There exists extensive literature on characterization of the distributed lag function and inference, assuming the constraints are correctly specified. Very few strategies exist for testing given distributed lag (DL) constraints. In this letter, we propose a new simple and efficient framework for testing a constrained DLM against an unconstrained DLM. In Section 2, we briefly introduce DLMs and present the proposed variance component score test (VCST) procedure. In Section 3, we conduct a simulation study to compare the statistical power of the standard likelihood ratio test (LRT) and VCST and illustrate both of the approaches using the National Mortality, Morbidity, and Air Pollution Study (NMMAPS) data. We conclude with discussions in Section 4.

* Corresponding author.
E-mail address: yinhsiuc@umich.edu (Y.-H. Chen).

http://dx.doi.org/10.1016/j.spl.2016.12.003

2. Method

Let x_t denote the independent variable measured at time t, y_t denote the response variable measured at time t, $\boldsymbol{z}_t$ denote the other covariates obtained at time t, T be the length of the time series, and L be the pre-determined maximum number of lags. Without loss of generality, we leave out intercept and covariates in the rest of the presentation and consider the generalized linear model $\eta_t = g[\mu_t] = g[E(y_t|x_t, x_{t-1}, \ldots, x_{t-L})] = \boldsymbol{X}_t^T\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_L)^T$ is the vector of the lag effects, $\boldsymbol{X_t} = (x_t, x_{t-1}, \ldots, x_{t-L})^T$, η_t is the canonical (natural) parameter, $g(\cdot)$ is the link function, y_t is a random variable generated from a distribution $\mathcal{F}$ in canonical exponential family with probability density

$$f(y_t) = \exp\left\{\frac{y_t\eta_t - b(\eta_t)}{a(\phi)} + c(y_t;\phi)\right\}, \tag{1}$$

ϕ is the dispersion parameter, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions. It is well-known that the exponential family possesses the properties of $\mu_t = b'(\boldsymbol{X}_t^T\boldsymbol{\beta}) = g^{-1}(\boldsymbol{X}_t^T\boldsymbol{\beta})$ and $V(y_t) = b''(\boldsymbol{X}_t^T\boldsymbol{\beta})a(\phi) = \nu(\boldsymbol{X}_t^T\boldsymbol{\beta})a(\phi)$ where $\nu(\cdot)$ is the variance function.

2.1. Constrained DLM

Constrained DLM imposes a pre-specified structure to constrain the lag coefficients to be a smooth function of the lags (i.e. $\beta_\ell = f(\ell)$ for $\ell = 0, \ldots, L$). Denote the p basis functions that generate the class of functions in which $\boldsymbol{\beta}$ can lie as $B_1(\cdot), \ldots, B_p(\cdot)$. The transformation matrix $\boldsymbol{C}$ as defined by Gasparrini et al. (2010) is given by

$$\boldsymbol{C} = \begin{bmatrix} B_1(0) & B_2(0) & \cdots & B_p(0) \\ B_1(1) & B_2(1) & \cdots & B_p(1) \\ \vdots & \vdots & \vdots & \vdots \\ B_1(L) & B_2(L) & \cdots & B_p(L) \end{bmatrix}_{(L+1)\times p}.$$

The constrained DLM estimator can be expressed in the form of $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\theta}$ where $\boldsymbol{\theta}$ is a vector of p free parameters to be estimated in $\mathbb{R}^p$. The maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T}\left[y_t\boldsymbol{X}_t^T\boldsymbol{C}\boldsymbol{\theta} - b(\boldsymbol{X}_t^T\boldsymbol{C}\boldsymbol{\theta})\right].$$

The estimation of $\boldsymbol{\theta}$ does not involve the dispersion parameter ϕ. The constrained DLM estimator is given by $\hat{\boldsymbol{\beta}}_{CDLM} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$. The $\boldsymbol{C}$ corresponding to an unconstrained DLM is a $(L+1)\times(L+1)$ identity matrix.

2.2. Hypothesis testing

$\hat{\boldsymbol{\beta}}_{CDLM}$ can be alternatively obtained by maximizing log-likelihood with respect to $\boldsymbol{\beta}$ subject to $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{R}$ is a $(L+1-p)\times(L+1)$ constraint matrix corresponding to the transformation matrix $\boldsymbol{C}$ (Chen et al., 2016). The basis functions in $\boldsymbol{C}$ span the solution space of $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$. $\boldsymbol{R}$ can be obtained from $\boldsymbol{C}$ via the following procedure. Define $\boldsymbol{C}_e$ as a $(L+1)\times(L+1)$ matrix $[\boldsymbol{C}\boldsymbol{0}_{(L+1)\times(L+1-p)}]$ where $\boldsymbol{0}_{(L+1)\times(L+1-p)}$ is a $(L+1)\times(L+1-p)$ matrix with zero entries. Applying singular value decomposition (SVD) $\boldsymbol{C}_e^T = \boldsymbol{U}_C\boldsymbol{D}_C\boldsymbol{V}_C^T$ where $\boldsymbol{U}_C$ is the $(L+1)\times(L+1)$ unitary matrix with left-singular column vectors, $\boldsymbol{V}_C$ is the $(L+1)\times(L+1)$ unitary matrix with right-singular column vectors, and $\boldsymbol{D}_C$ is a $(L+1)\times(L+1)$ diagonal matrix with singular values of $\boldsymbol{C}_e^T$ along the diagonal, the constraint matrix $\boldsymbol{R}$ can then be obtained as the last $(L+1-p)$ rows of $\boldsymbol{V}_C^T$.

Testing a particular DLM structure against an unconstrained alternative can now be formulated as testing $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{0}$. A standard likelihood ratio test (LRT), Wald test, and score test can be conducted and the test statistics asymptotically follow a χ^2 distribution with $L+1-p$ degrees of freedom and large sample inference can be obtained. We propose a VCST approach to this problem. Consider a generalized ridge regression estimator (Chen et al., 2016) that minimizes the penalized negative log-likelihood function

$$\ell_p(\boldsymbol{\beta}) = \left[S(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta}\right] + \lambda\boldsymbol{\beta}^T\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_T)^T$, $\boldsymbol{X}$ is a $T\times(L+1)$ matrix with $\boldsymbol{X}_t^T$ as the tth row for $t = 1, \ldots, T$, $S(\cdot)$ is the $\mathbb{R}^T \to \mathbb{R}^1$ cumulant function such that $S(\boldsymbol{a}) = \sum_{t=1}^{T} b(a_t)$ with $\boldsymbol{a} = (a_1, \ldots, a_T)^T$, and λ is the tuning parameter. We can rewrite

$$\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T \tag{3}$$

where $\boldsymbol{U}$ is a $(L+1)\times(L+1)$ matrix with orthogonal columns and $\boldsymbol{D}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{R}^T\boldsymbol{R}$ using singular value decomposition (SVD). Since $\boldsymbol{R}^T\boldsymbol{R}$ is not of full rank and has rank $L+1-p$, we can write $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$ where $\boldsymbol{D}_1$ is a $(L+1-p)\times(L+1-p)$ diagonal matrix of full rank. Let $\boldsymbol{U} = [\boldsymbol{U}_1\boldsymbol{U}_2]$ where $\boldsymbol{U}_1$ is a $(L+1)\times(L+1-p)$ matrix