

Fitting mean-shift models with covariates

Stima di modelli con sfasamento del valor medio in presenza di variabili esplicative

Anna Maria Taormina, Vito M.R. Muggeo, Giancarlo Ferrara

Dipartimento di Scienze Statistiche e Matematiche “S. Vianelli”, Università di Palermo
e-mail: taormina@dssm.unipa.it

Riassunto: In questo lavoro viene proposto un semplice algoritmo di stima dei punti di svolta in modelli con punti-di-cambio e sfasamenti nel valor medio in presenza di variabili esplicative con coefficienti fissi, cioè non variabili nel tempo. Il metodo lavora con qualsiasi numero di *break-point* e si basa su una semplice trasformazione dei residui. Uno studio di simulazione ed un’analisi di un dataset vengono presentati per illustrare gli aspetti pratici dell’algoritmo.

Keywords: change-point, mean-shift model, parameter instability, segmented regression

1. Introduction

In econometric time series regression parameter instability is a frequent peculiarity, in particular for data covering an extended period and subjected to structural changes. Roughly speaking, models which allow abrupt changes in parameters are called ‘break-points models’ and when the change concerns just the mean level, they are more specifically called ‘mean-shift models’ (Bai and Perron, 2003). Among the different aspect of interest, this paper deals with model estimation when i) there are m unknown break-points and ii) the mean level of the response depends on a set of covariates $\mathbf{x}_t$ whose coefficient are fixed-in-time. Formally for $t = 1, 2, \dots, T$ observations, we consider the following model:

$$y_t = \alpha_0 + \sum_{i=1}^m \alpha_i I(t > \psi_i) + \beta' \mathbf{x}_t + u_t \quad t = 1, \dots, T, \quad (1)$$

being $I(a) = 1$ if a is true, $\mathbf{x}_t$ is the covariate vector observed at time t with coefficients β , ψ_i is the i^{th} break-point, i.e. the time point where the mean level α_{i-1} rises to $(\alpha_{i-1} + \alpha_i)$, ($i = 1, \dots, m$) and u_t is the usual white noise, i.e. $u_t \sim \mathcal{N}(0, \sigma^2)$.

Structural change and parameter instability have been receiving attention in the econometrics and statistics communities with particular emphasis on model estimation, specially in presence of multiple breakpoints; although some results have been attained, these are computationally intensive and thus are not favored in practice. Among others, Sullivan (2002) criticizes the method that locates multiple shifts by recursively dividing the data and proposes an algorithm for simultaneously estimating the locations of the changes based on the application of the Schwarz Information Criterion; Bai and Perron (2003) present an efficient algorithm based on dynamic programming which requires at most operations of order $O(T^2)$ for any number of breaks. In this paper we propose an alternative method that is more efficient than those mentioned, especially with $m > 1$.

The paper is structured as follows: in section 2 the estimating algorithm is described and in section 3 we present results from a simulation study and from a real dataset analysis; section 4 is devoted to conclusions.

2. Methods

In this section we describe a method for an efficient estimation of a mean-shift model with m breaks. The proposed procedure is based on an exact yet efficient algorithm designed for estimation of regression models with *continuous* break-points, i.e. models where the regression lines themselves are connected (Muggeo, 2003). The idea is fairly simple: since such method can be successfully used only with continuous breakpoint, we suggest to transform data to making use of such algorithm even in presence of jump-point, namely the mean-shift model (1).

The algorithm is simple and consists of the following three steps:

1. fit the linear model $E(y_t) = \alpha + \beta'x_t$;
2. take the residuals e_t , say, compute the cumulative sums $r_t = \sum^t e_k$ in order to obtain a continuous piecewise relationship with the time-index variable $t = 1, 2, \dots, T$ and apply the algorithm proposed in Muggeo(2003) for the estimation of m break-points in the continuous breakpoint model $E[r_t] = \delta_{00} + \delta_0 t + \sum_{i=1}^m \delta_i(t - \psi_i)_+$; let $\hat{\psi}_i$ the point estimates;
3. use the estimated break-points gained in 2., and fit the final mean-shift model:

$$E(y_t) = \alpha_0 + \sum_{i=1}^m \alpha_i I(t > \hat{\psi}_i) + \beta'x_t.$$

Therefore estimation is performed in two steps: first breakpoint estimates are obtained in 2. using residuals from step 1. and then the final model is estimated in 3. The ‘cumulative sum’ is the transformation employed to convert the (discontinuous) jump-point relationships into (continuous) breakpoints ones. Also, note that separate estimation is permitted because of the $\hat{\psi}_i$ s turn out to be uncorrelated with the estimators of the other parameters (Bai and Perron, 2002).

Final model estimation and inference may be accomplished using previous results based on point estimates. Following Bai (1997), confidence interval for a single break-point can be computed, namely: given $\hat{L}_i = \hat{\alpha}_i^2 / \hat{\sigma}^2$, a $100(1 - \gamma)\%$ confidence interval has extremes $[\hat{\psi}_i \pm c_\gamma / \hat{L}_i]$, where c_γ is the $(1 - \gamma/2)$ th quantile of a particular random variable (Bai, 1997); hypothesis testing comparing models with m versus $m + 1$ breakpoints is also straightforward following formulas found in Bai and Perron (2003).

3. Simulation study

A simulation study has been carried out in order to assess the validity of the method proposed; for sake of simplicity only a single breakpoint is taken into account. Different scenarios have been considered: different sample sizes ($T = \{50, 200\}$), different break-point locations ($\psi = \{T/2, 3T/4\}$) and different mean-shifts values ($\alpha_1 = \{1.5, 3\}$). Given T , ψ and α_1 , the model being analyzed is:

$$y_t = 1 + \alpha_1 I(z_t > \psi) + 5x_{1t} + 0.5x_{2t} + u_t \quad t = 1, \dots, T,$$

where the covariates $x_{1t} \sim U(0, 1)$ and $x_{2t} \sim \log \mathcal{N}(3, 1)$ and $u_t \sim \mathcal{N}(0, 9)$. Hereafter we use $z_t = t/T$ instead of t with ψ scaled accordingly.

Table 1 summaries the performance of the break-point estimator: Montecarlo estimates (based on 1000 replicates) of the mean, median and standard deviation (SD) have been reported for each combination of T , ψ and α_1 . Furthermore, in order to assess the adequacy of the formula proposed by Bay and Perron (2003) to compute confidence interval for the breakpoint, we have also reported the relevant coverage probability.

Table 1: *Simulation results for break-point estimator (1000 replicates) in a mean shift-model: mean, median, standard deviation of ψ and coverage probability of the estimated 95% CI.*

Model	T	ψ	Sampling distribution summary			%CP
			Mean	Median	SD	95 per cent CI
$\alpha_1 = 1.5$	50	0.50	0.477	0.480	0.130	97.0
		0.75	0.694	0.720	0.162	92.5
	200	0.50	0.493	0.495	0.079	97.6
		0.75	0.729	0.745	0.099	94.4
$\alpha_1 = 3$	50	0.50	0.493	0.500	0.088	95.5
		0.75	0.717	0.740	0.113	93.6
	200	0.50	0.497	0.495	0.032	94.6
		0.75	0.742	0.745	0.037	95.5

The values reported show that bias and variance of the break-point estimator obtained via the method described in section 2 yield reasonable values which reduce when sample size increases; moreover, as one could expect it, location in the middle and a larger shift importantly influence the performance of the estimator.

4. An application

We apply the method discussed in section 2 to a time series giving the monthly totals of car drivers in Great Britain killed or seriously injured from Jan 1969 to Dec 1984 (Harley and Durbin, 1986); we choose this dataset because of a well-grounded suspicion of structural changes in the mean level. Following Harley and Durbin we use the logarithm of the data and regress it on lagged values at lag 1 and 12. To apply our method, we first fit a linear model for the log-trasformed values (y) versus the two explanatory variables y_{lag1} and y_{lag12} and then we use the relevant cumulative residuals to estimate the breakpoint location. The final two break-point fitted model with residual sum of squares (RSS) equal to 0.300 and $AIC = -628.36$ is

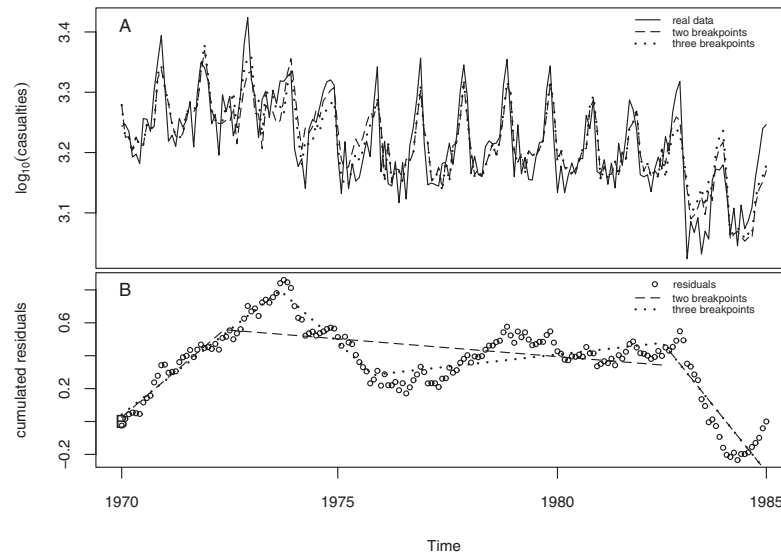
$$E(y_t) = 0.54 - 0.02I(z_t > 0.16) - 0.03I(z_t > 0.87) + 0.33y_{lag1_t} + 0.51y_{lag12_t}.$$

A three break-points fitted model with $RSS = 0.286$ and $AIC = -635.53$ is:

$$E(y_t) = 0.42 - 0.04I(z_t > 0.25) + 0.03I(z_t > 0.39) - 0.02I(z_t > 0.84) + 0.31y_{lag1_t} + 0.57y_{lag12_t}.$$

Figure 1 summarizes results: panel A displays the observed data (log-transformed values) along with the expected values from the two fitted models, while panel B shows the cumulative residuals with the relevant “segmented” fitted values: the time points where the straight lines join correspond to the estimated break-points.

Figure 1: Log-transformed time series with the two fitted mean-shift models (A); cumulated residuals of the linear model with the “segmented ” fitted values (B)



5. Discussion

A simple algorithm to fit mean shift model has been illustrated in this paper. The method is particularly efficient with any number of breaks and it works also when explanatory variables with fixed coefficients have to be considered. A few simulations have shown that the resulting estimator is approximatively unbiased with variance decreasing as sample size increases. Although we have dealt only with changes in the intercept parameter, future research will be addressed to extend the idea to general change-point models where any regression parameter is allowed to change in time.

References

- Bai, J. , Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18, 1–22.
- Harvey, A. C. and Durbin, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society series B*, 149, 187–227.
- Muggeo, V. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22, 3055–3071.
- Sullivan, J. H. (2002) Estimating the location of multiple change points in the mean. *Computational Statistics*, 17, 289–296.