Automatic covariates selection in dynamic regression models with application to COVID-19 evolution

Ana X. Ezquerro[1], Germán Aneiros Pérez[2], and Manuel Oviedo de la Fuente[3]

[1] University of A Coruña
ana.ezquerro@udc.es
[2] Departamento de Matemáticas, University of A Coruña
german.aneiros@udc.es
[3] Departamento de Matemáticas, University of A Coruña
manuel.oviedo@udc.es

Abstract

This work introduces a new approach in time-series analysis field for automatic covariates selection in dynamic regression models. Based on [2] and [5] previous study, a forward-selection method is proposed for adding new significant covariates from a given set. This algorithm has been implemented and optimized in R as a package, and it has been applied to multiple simulations to validate its performance. Finally, the obtained results from the IRAS database of Catalonia are presented to analyze the COVID-19 evolution.

Contents

1 Introduction

In time-series analysis, the well-known dynamic regression models allow formally modelling the dependence between a set of covariates and a dependent variable considering the intrinsic temporal component of all participant variables. Thus, this type of regression models are of widespread application in diverse scenarios where it is required to analyze the effect of recollected data in a time series of interest.

Formally, dynamic linear regression models define the lineal dependence between a stochastic proccess Y_t (the dependent variable) and a set of proccesses $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(m)}\}$ (candidates for regressor variables) in times non-greater than t:

$$Y_t = \beta_0 + \beta_1 X_{t-r_1}^{(1)} + \beta_2 X_{t-r_2}^{(2)} + \cdots + \beta_m X_{t-r_m}^{(m)} + \eta_t \tag{1}$$

where $r_i \geq 0$, for $i = 1, ..., m$, and $\eta_t \sim$ ARMA(p,q).

In this work we formally introduce a new algorithm to select covariates which significantly influence the behavior of a dependent variable. Due to the impact of COVID-19 around the world, we use this method to formalize and study the relation of the COVID-19 evolution in Catalonia (Spain) with the flu syndrome, COVID-19 vaccination and other recollected variables from the IRAS database.

2 Methodology

Following the definition in 1, [2] proposed a method named *prewhitening* for removing spurious correlation (false lineal correlation) between two processes X_t and Y_t (where non of them were white noise) by analyzing the cross correlation function

$$\rho_k(X_t, Y_t) = \frac{\text{Cov}(X_t, Y_{t-k})}{\sigma_{X_t}\sigma_{Y_t}} \tag{2}$$

where σ_{Z_t} denotes the standard deviation of a stochastic process Z_t. Specifically, [2] proposes a real lineal correlation between X_t and Y_t if exists some k where $\rho_k(X_t, Y_t)$ is statistically significant. This method is applied to obtain the optimal lags of each regressor in 1, considering the condition of k being less or equal than 0.

Our approach iteratively adds dependent processes to a model by checking if a significant correlation (analyzing 2 as in [2]) exists between a new proccess (candidate for regressor variable) and the residuals η_t of a simpler model.

Let Y_t be the stochastic dependent proccess and $\mathcal{X}$ be the set of processes that might act as regressor variables in the model (candidates), and an information criterion for model evaluation. Our method proceeds as follows:

1. Initialization. Consider the process $\tilde{Y}_t = Y_t$ that will be used to check the existence of correlation between the proccesses in $\mathcal{X}$ with [2] method, $\nu = \infty$ the value of the best model with 1 form and $\mathcal{X}_{\mathcal{M}} = \emptyset$ the set of selected covariates (initially empty) with their respective optimal lags and coefficients.

2. Iterative selection. For each $X_t \in \mathcal{X} - \mathcal{X}_{\mathcal{M}}$, obtain the optimal lag where the maximum linear cross correlation between X_t and $\tilde{Y}_t$ occurs in (via [2] method). Consider the$X_t \in \mathcal{X} - \mathcal{X}_{\mathcal{M}}$ that improves ν value, based on the selected information criterion:

$$X_t^{\text{best}} = \arg\min_{X \in \mathcal{X}} \left\{ \text{criteria}(\tilde{\mathcal{M}}) \right\}$$

where

$$\mathcal{M} = Y_t = \beta_0 + \beta_{\text{best}} X_{t-r_{\text{best}}} + \sum_{(X_t, r, \beta) \in \mathcal{X}_{\mathcal{M}}} \beta X_{t-r} + \eta_t$$

conditioned to criteria$(\mathcal{M}) < \nu$. If X_t^{best} exists, consider $\mathcal{X}_{\mathcal{M}} = \mathcal{X}_{\mathcal{M}} \cup \{X_t^{\text{best}}\}$, $\tilde{Y}_t = \eta_t$ and $\nu = $ criteria$(\mathcal{M})$. Repeat this step until no process $X_t \in \mathcal{X} - \mathcal{X}_{\mathcal{M}}$ can be added to $\mathcal{X}_{\mathcal{M}}$.

3. Finalization. If the errors η_t of $\mathcal{M}$ are not stationary and no model with $\eta_t \sim$ ARMA(p,q) can be adjusted, consider the regular differentiation of all data (dependent variable and regressor candidates) and return to (1). Otherwise, it is proven that $\mathcal{M}$ with stationary errors defines the significant correlation between the set of $\mathcal{X}_{\mathcal{M}}$ regressor variables and the dependent proccess Y_t.

This algorithm was implemented in R programming language. The step (2) was optimized by parallelizing the fit of independent models of each candidate in $\mathcal{X}$. [3] test is used for checking proccesses stationary and [4, 8, 6, 1, 7] tests to check the independence, normality and zero mean of ARIMA residuals.

3 Simulation results

In order to evaluate the performance of our selection method, we simulate multiple scenarios where a time series Y_t was artificially constructed with other variables (introduced with their respective coefficients and lags as in 1), which were added to a set of candidates along with more variables which do not influence in the construction of Y_t. The algorithm was tested when the residuals of the model η_t were stationary and non-stationary.

4 COVID-19 application

References

[1] G. E. P. Box and David A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526, 1970.

[2] Jonathan D Cryer and Kung-Sik Chan. *Time series analysis: with applications in R*, chapter 11. 2, 2008.

[3] David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

[4] Winston Haynes. *Student's t-Test*, pages 2023–2025. Springer New York, New York, NY, 2013.

[5] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*, chapter 9. OTexts, 2018.

[6] Carlos M. Jarque. *Jarque-Bera Test*, pages 701–702. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[7] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 08 1978.

[8] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.