IX Xornada de Usuarios de R en Galicia
Santiago de Compostela, 20 de outubro do 2022

New covariates selection method in dynamic regression models with a public implementation in R language

Ana Ezquerro[1], Germán Aneiros[2], Manuel Oviedo[3]

[1]University of A Coruña, `ana.ezquerro@udc.es`
[2]CITIC, Grupo MODES, Departamento de Matemáticas, University of A Coruña, `german.aneiros@udc.es`
[3]CITIC, Grupo MODES, 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 [1] and [2] 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 we openly publish its sources in order to make it available for all the R community. Our method 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.

Keywords: Time series, dynamic regression models, selection methods, forecasting.

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 desired to analyze the effect of recollected data in a time series of interest.

Formally, dynamic linear regression models define the linear dependence between a stochastic process Y_t (the dependent variable) and a set of processes $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(m)}\}$ (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. The implementation of this selection method is publicly available[1].

Methodology

Following the definition in 1, [1] proposed a method named *prewhitening* for removing spurious correlation (false linear correlation) between two processes X_t and Y_t (where one of them is not white noise and/or the other is not stationary) by analyzing the cross correlation function $\rho_k(\ddot{X}_t, \ddot{Y}_t) = \dfrac{\text{Cov}(\ddot{X}_t, \ddot{Y}_{t-k})}{\sigma_{\ddot{X}_t}\sigma_{\ddot{Y}_t}}$ where σ_{Z_t} denotes the standard deviation of a stochastic process Z_t

[1]`https://github.com/anaezquerro/dynamic-arimax`

and $\ddot{X}_t$ and $\ddot{Y}_t$ are obtained via some linear filter application to X_t and Y_t ensuring one of them is white noise and the other is a stationary process. Specifically, [1] proposes a real linear correlation between X_t and Y_t if exists some k where $\rho_k(\ddot{X}_t, \ddot{Y}_t)$ 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 (as in [1]) exists between a new process (candidate for regressor variable) and the residuals η_t of a simpler model.

Let Y_t be the stochastic dependent process and $\mathcal{X}$ be the set of processes that might act as regressor variables in the model (candidates), and an information criterion (IC) 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 linear correlation between Y_t and each $X_t \in \mathcal{X}$ with [1] method, $\nu = \infty$ the value of the IC corresponding to the best model with 1 form, $\mathcal{X}^{(s,r)}$ the set of selected covariates paired with their respective optimal lags and $\mathcal{X}^{(s)}$ the set of selected covariates (with no lag information). Let $\mathcal{M}(\mathcal{Z})$ be the fitted dynamic regression model regarding Y_t where $\mathcal{Z}$ is the set of covariates paired with their optimal lags:

   $$\mathcal{M}(\mathcal{Z}) := Y_t = \beta_0 + \sum_{(Z_t,r)\in\mathcal{Z}} \beta^{(Z_t,r)} Z_{t-r} + \eta_t$$

   where $\beta^{(Z_t,r)}$ is obtained via some estimation.

2. Iterative selection. For each $X_t \in \mathcal{X} - \mathcal{X}^{(s)}$, obtain the optimal lag where the maximum linear cross correlation between X_t and $\tilde{Y}_t$ occurs (via [1] method). Consider the process $X_t^{\text{best}} \in \mathcal{X} - \mathcal{X}^{(s)}$ that minimizes and improves ν value, based on the selected IC, by including it in the model with its optimal lag ($r_{X_t^{\text{best}}}$):

   $$X_t^{\text{best}} = \underset{X_t\in\mathcal{X}-\mathcal{X}^{(s)}}{\arg\min} \left\{ \text{criteria}\left(\mathcal{M}\left(\mathcal{X}^{(s,r)} \cup \left\{(X_t, r_{X_t})\right\}\right) \right) \right\} \tag{2}$$

   conditioned to criteria$(\cdot) < \nu$[2]. If X_t^{best} exists, consider $\mathcal{X}^{(s,r)} = \mathcal{X}^{(s,r)} \cup \{(X_t^{\text{best}}, r_{X_t^{\text{best}}})\}$, $\tilde{Y}_t = \eta_t$ and $\nu = \text{criteria}(\mathcal{X}^{(s,r)})$[3]. Repeat this step until no process $X_t \in \mathcal{X} - \mathcal{X}^{(s)}$ can be added to the model, i.e. X_t^{best} does not exist.

3. Finalization. If the errors η_t of $\mathcal{M}(\mathcal{X}^{(s,r)})$ are not stationary and no model with $\eta_t \sim$ ARMA(p,q) and $\mathcal{X}^{(s,r)}$ covariates can be adjusted, consider the regular differentiation of all data (dependent variable and regressor candidates) and return to (1). Otherwise, it is proven that $\mathcal{M}(\mathcal{X}^{(s,r)})$ with stationary errors defines the significant correlation between the set of $\mathcal{X}^{(s)}$ regressor variables and the dependent process 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 $\mathcal{X}$. Dickey-Fuller test is used for checking processes stationary, Ljung-Box to check the independence, Shapiro-Wilks and Jarque-Bera tests for normality and t-test for zero mean of ARIMA residuals.

Simulation results

In order to validate 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.

Specifically, we simulate $M = 100$ times the following scenario:

[2] for simplicity, we denote the expression in criteria() in 2 as $\cdot$

[3] once X_t^{best} has been added to the model

Figure 1: Example of code output and results of `drm.select()` when running the selection method

```
beta0 <- -0.6; beta1 <- 1.7; beta2 <- -2.2; beta3 <- 1.3; r1 <- 2; r3 <- 3
Y <- beta0 + beta1*lag(X1,-r1) + beta2*X2 + beta3*lag(X3,-r3) + residuals
xregs <- cbind(X1, X2, X3, X4, X5, X6)
ajuste <- drm.select(Y, xregs, ic='aicc', st_method='adf.test', show_info=F)

print(ajuste$history, row.names=F)

 var lag                ic
  X2   0 -1156.68486061937
  X1  -2 -2171.66958134745
  X3  -3 -3108.15443209894

print(ajuste, row.names=F)

Series: serie
Regression with ARIMA(0,0,4) errors

Coefficients:
         ma1     ma2  ma3     ma4  intercept       X2      X1      X3
      0.2498  0.3360    0  0.1589    -0.5947  -2.1868  1.6949  1.3083
s.e.  0.0304  0.0302    0  0.0300     0.0033   0.0105  0.0089  0.0320

sigma^2 = 0.002377:  log likelihood = 1562.15
AIC=-3108.3   AICc=-3108.15   BIC=-3069.26
```

1. We generate seven different independent time series (each modelable by an ARIMA), of which six of them act as the covariate candidates set: $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(6)}\}$; and the remaining as the residuals η_t of the model.

2. We construct the dependent variable Y_t by a linear combination of $\{X_t^{(1)}, X_t^{(2)}, X_t^{(3)}\}$, randomly lagged $r = 0, ..., 6$ moments (where the coefficients are randomly generated), with an intercept β_0 and the generated residuals η_t. Formally,

$$Y_t = \beta_0 + \beta_1 X_{t-r_1}^{(1)} + \beta_2 X_{t-r_2}^{(2)} + \beta_3 X_{t-r_3}^{(3)} + \eta_t$$

   where $\beta_0, ...\beta_3$ are randomly generated, and $r_i \in [0, 6]$ for $i = 1, 2, 3$.

3. We launch our selection method with different configurations:
   - Using as the information criterion the AIC, BIC and AICc.
   - Using as the method to check stationary the Dickey-Fuller test or via analyzing the differentiation order when an ARIMA is adjusted.

4. Evaluate the selection method using as metrics the percentage of times a covariate is:
   (a) correctly added to the model (*true positive*),
   (b) incorrectly added to the model (*false positive*),
   (c) correctly not added to the model (*true negative*),
   (d) incorrectly not added to the model (*false negative*).

Figure 1 displays the result of calling the function that implements the selection method. The DataFrame stored in `$history` provide information about the covariates iteratively added to the model, the IC value achieved and the lag they were added with. When printing the resultant object (`ajuste`) we see that the estimated regression coefficients are nearly the same than the real values artificially set. Also, the lags estimated by the method are correct, and the errors of the final model are stationary.

Table illustrates a resume of the results of our approach when running our approach $M = 100$ times with different configurations and using stationary an non-stationary errors.

Table 1: Percentage data results with different configurations when residuals are stationary

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	97.66%	97.66%	97.66%	3.66%	1.33%	3.66%
auto.arima	98.33%	98.33%	98.33%	3.66%	1.33%	3.66%
	correctly added (TP)			incorrectly added (FP)		

	AIC	**BIC**	**AICc**	**AIC**	**BIC**	**AICc**
adf.test	96.33%	98.66%	96.33%	2.33%	2.33%	2.33%
auto.arima	96.33%	98.66%	96.33%	1.66%	1.66%	1.66%
	correctly **not** added (TN)			incorrectly **not** added (FN)		

Table 2: Information about the dynamic regression model constructed via selection of multiple vaccination variables to model COVID19 evolution

Covariate	**Lag**	**Coefficient est. (s.e)**
`vac4565`	-3	-0.0410 (0.0057)
`vac6580`	-2	-0.0468 (0.0120)
`vac1845`	-6	-0.0901 (0.0047)
`vac1218`	Not included in the model	
`vac80`	Not included in the model	
residuals	ARIMA(4, 0, 0)	$\phi_1 = 2.0816(0.0810)$ $\phi_2 = -1.2837(0.1152)$ $\phi_4 = 0.1919(0.0432)$

Application to COVID19 evolution

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. Individual data was aggregated by age ranges and Health Areas to study the correlation between groups and their influence in the global evolution.

Table 2 resumes the algorithm trace and the order of covariates addition to the model. The covariates named `vac1218`, `vac1845`, `vac4565`, `vac6580` correspond the vaccination data in population from 12, 18, 45 and 65 up to 18, 45, 65 and 80 years old (exclusive), and `vac80` corresponds the vaccination in population from 80 years. We can analyze the vaccination has a negative impact in the expansion of COVID19, specifically, the vaccination of working-age population.

Future work

Our approach has considered DRM covariates modelable by ARIMA models, which successfully covers a wide real-life applications. However, other cases might be considered, such as adding functional variables and discrete variables to the set of candidates.

Acknowledgements

To Banco Santander for the scholarships offered in 2021/2022, which helped the investigation of this proposal, and to *Maths Department* of University of A Coruña.

References

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

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