Selección automática de variables regresoras en modelos de regresión lineal dinámica con aplicaciones al COVID-19

Ana Xiangning Pereira Ezquerro
ana.ezquerro@udc.es

Germán Aneiros Pérez
german.aneiros@udc.es

Manuel Oviedo de la Fuente
manuel.oviedo@udc.es

Departamento de Matemáticas, Curso 2021-2022

En el estudio y análisis de series temporales, los bien conocidos modelos de regresión dinámica permiten modelar formalmente la dependencia entre un conjunto de variables regresoras y una variable dependiente teniendo en cuenta la componente temporal inherente a todas las variables participantes. En base al trabajo previo de [Cryer and Chan, 2008] y [Hyndman and Athanasopoulos, 2018], se propone un método iterativo de tipo *forward selection* para añadir variables significativas de un conjunto dado al modelo de regresión con sus respectivos retardos óptimos.

Formalmente, los modelos de regresión lineal dinámica expresan la dependencia lineal entre un proceso Y_t y un conjunto de procesos $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(m)}\}$ en instantes no superiores a t:

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

donde $r_i \geq 0$, para $i = 1, ..., m$, y $\eta_t \sim$ ARMA(p,q).

[Cryer and Chan, 2008] propuso un método denominado *preblanqueado* para eliminar la correlación espuria entre dos procesos X_t e Y_t y así detectar de forma limpia la existencia de dependencia lineal entre ellos, así como el retardo r en el que se produce dicha dependencia. Siguiendo su metodología, esta nueva aproximación permite añadir de forma iterativa procesos dependientes al modelo en base a chequear si existe una correlación significativa (siguiendo el método de [Cryer and Chan, 2008]) entre un nuevo proceso y los residuos η_t del modelo.

Dado un proceso Y_t y un conjunto de procesos estocásticos $\mathcal{X} = \{X_t^{(1)}, X_t^{(2)}, ..., X_t^{(m)}\}$ que actúan como variables regresoras en el modelo, y un criterio de información para evaluar el modelo, el procedimiento a seguir es:

1. Buscar en todo el conjunto $\mathcal{X}$ el proceso con su retardo óptimo que mejor modelo de regresión lineal dinámica simple produzca, en base al criterio de información escogido.

$$X_t^{\text{best}} = \underset{X \in \mathcal{X}}{\text{arg mín}} \Big\{ \text{criteria}\big(Y_t = \beta_0 + \beta_1 X_{t-r} + \eta_t\big) \Big\}$$

2. Si X_t^{best} existe, construir el modelo $\mathcal{M} : Y_t = \beta_0 + \beta_1 X_{t-r}^{\text{best}} + \eta_t$, retirar X^{best} de $\mathcal{X}$, tomar $\tilde{Y}_t = \eta_t$ y proceder de forma iterativa:

   2.1. Buscar en $\mathcal{X}$ el proceso X_t^{best} tal que:

$$X_t^{\text{best}} = \underset{X \in \mathcal{X}}{\text{arg mín}} \Big\{ \text{criteria}\big(\tilde{Y}_t = \beta_0 + \beta_1 X_{t-r} + \eta_t\big) \Big\}$$

   con la condición de que $\text{criteria}\big(\tilde{Y}_t = \beta_0 + \beta_1 X_{t-r} + \eta_t\big) < \text{criteria}(\mathcal{M})$

   2.2. En caso de que se pueda obtener dicho modelo, considerar un nuevo $\mathcal{M} : \tilde{Y}_t = \beta_0 + \beta_1 X_t^{\text{best}} + \eta_t$ y $\tilde{Y}_t = \eta_t$ y volver a (2.1). En otro caso, detener el algoritmo.

3. El modelo de regresión lineal dinámica se construye a partir de Y_t y el conjunto de procesos escogidos como X_t^{best} en cada iteración del algoritmo.

Esta nueva propuesta ha sido implementada y optimizada en el lenguaje estadístico R en formato de paquete, y se ha aplicado sobre una serie de simulaciones para validar su rendimiento. Finalmente se presentan los resultados obtenidos usando la base de datos de IRAS de Cataluña para analizar la evolución de la COVID-19.

Referencias

[Cryer and Chan, 2008] Cryer, J. D. and Chan, K.-S. (2008). *Time series analysis: with applications in R*, chapter 11. 2.

[Hyndman and Athanasopoulos, 2018] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*, chapter 9. OTexts.