Análisis de intervención y tratamiento de los puntos anómalos

Curso de Series Cronológicas año 2020

Silvia Rodríguez Collazo

Licenciatrua de Estadística
Facultad de Ciencias Económicas y de Administración

Las series se ven con frecuencia afectadas por **sucesos puntuales conocidos** . Ej:una huelga, una año bisiesto, un cambio legal o un cambio en un feriado. Si incluimos estos efectos en la serie podemos mejorar la precisión de la estimación de los parámetros y de las predicciones.

En una serie de producción conocemos que se ha producido una huelga en un momento t. Se puede incluir ese efecto a través de una variable indicadora que llamaremos variable impulso.Esa variable impulso, toma el valor 1 en el momento de la huelga y cero en el resto.

$$I_t = \begin{cases} 1 & \text{t} = \text{huelga} \\ 0 & \text{resto} \end{cases}$$

$$Z_t = \mu + wI_t + h_t$$

h_t tiene estructura temporal la que se representa un modelo ARIMA. (1)
Si estudiamos la relación entre los valores de la serie y la variable impulso se puede estimar el efecto de la intervención en la serie, representada por la variable impulso.

Veamos otro ejemplo, si se produce un cambio legal que modifica la definición de desempleo a partir de un determinado momento, este cambio se puede modelizar mediante una intervención de este tipo:

$$S_t = \begin{cases} 0 & t < \text{cambio legal} \\ 1 & t \geq \text{cambio legal} \end{cases}$$

Si S_t es una variable escalón, que toma el valor cero antes del cambio legal y uno luego de la ocurrencia del mismo.

$$Z_t = \mu + wS_t + h_t \quad h_t \quad \text{es un modelo ARIMA} \tag{2}$$

Representaremos estos sucesos mediante diversos tipos de variables.

Box y Tiao(1975) denominaron a ese análisis como análisis de intervención.La esencia del análisis de intervención es aislar los efectos de la intervención de otros acontecimientos y del resto de las perturbaciones aleatorias.

Las variables ficticias más utilizadas para representar estos sucesos cualitativos que afectan a la serie son de dos tipos, variables impulso y variables escalón.

Supongamos que Y_t sigue un modelo ARMA que podemos representar como:

$$Y_t = \Psi(L)\varepsilon_t \tag{3}$$

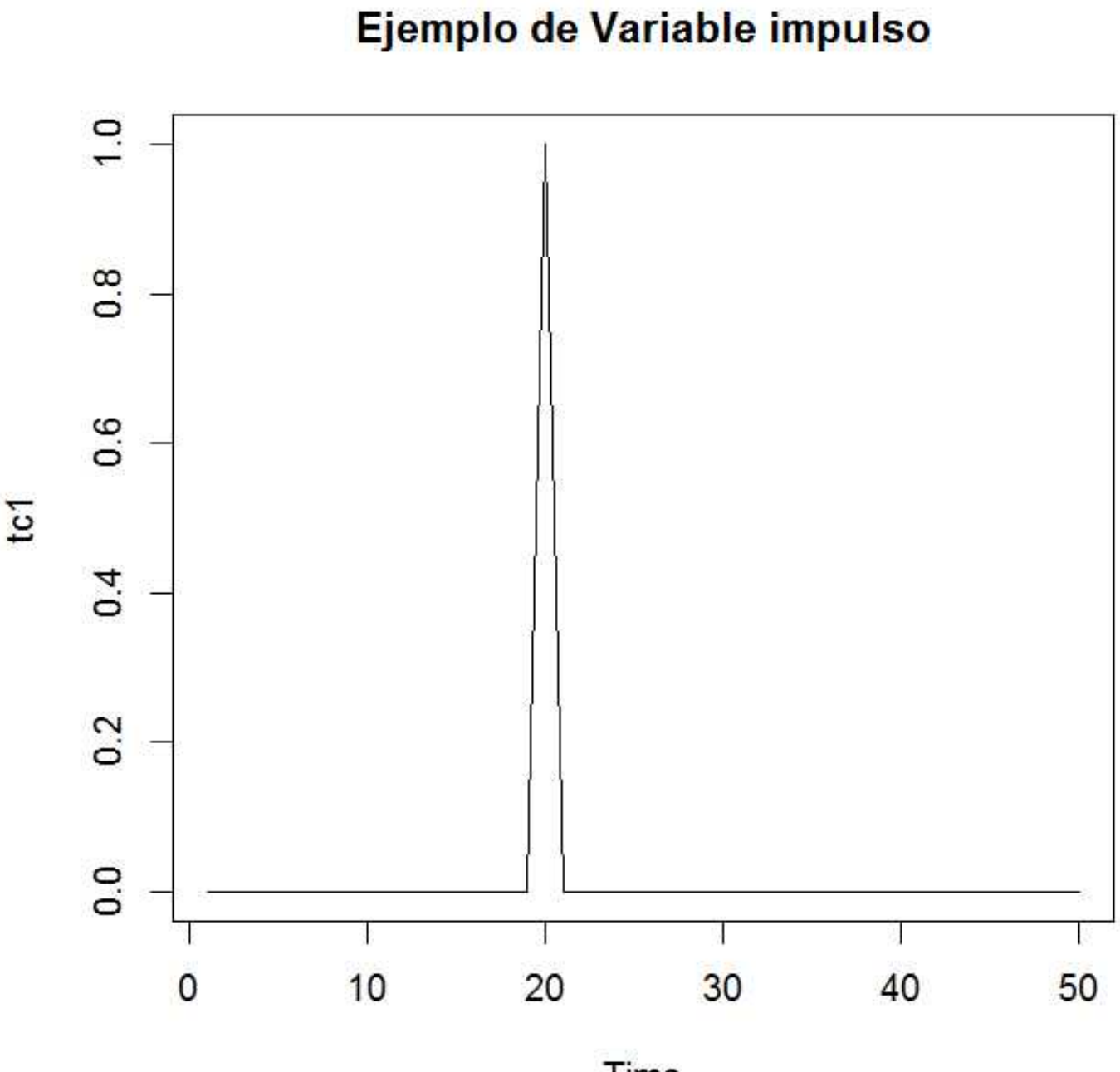
y esta serie está afectada en un momento conocido $t = h$ por un suceso también conocido.
Por lo que en el momento $t = h$ ya no observamos sólo a la variable Y_t sino también a Z_t que esta relacionada con Y_t.

$$Z_h = W_0 + Y_h \tag{4}$$

con W_0 magnitud del efecto sobre la serie.
Para representar el suceso definimos una **variable impulso**

$$I_t^h = \begin{cases} 0 & t \neq h \\ 1 & t = h \end{cases}$$

Variables Impulso. Función de repuesta a impulsos I



Variables Impulso. Función de repuesta a impulsos II

Para representar el efecto del suceso sobre la serie :

$$Z_t = W_0 I_t^h + y_t$$

y suponemos $Y_t = \Psi(L)\varepsilon_t$

$$Z_t = W_0 I_t^h + \Psi(L)\varepsilon_t \tag{5}$$

De este modo, la serie observada Z_t sigue el modelo ARMA $Z_t = \Psi(L)\varepsilon_t$ en todos los momentos del tiempo donde la variable impulso es cero, pero en $t = h$ se observa el valor de la suma de la realización del modelo ARMA más un efecto determinístico de tamaño W_0.

Pero el **efecto** de la intervención definida por el impulso I_t^h puede ser más complejo o duradero y distribuirse en varios períodos.

Consideremos el siguiente ejemplo: una lluvia torrencial en $t = h$, veamos el efecto sobre el tráfico, que puede durar n períodos, por tanto la serie que se observa:

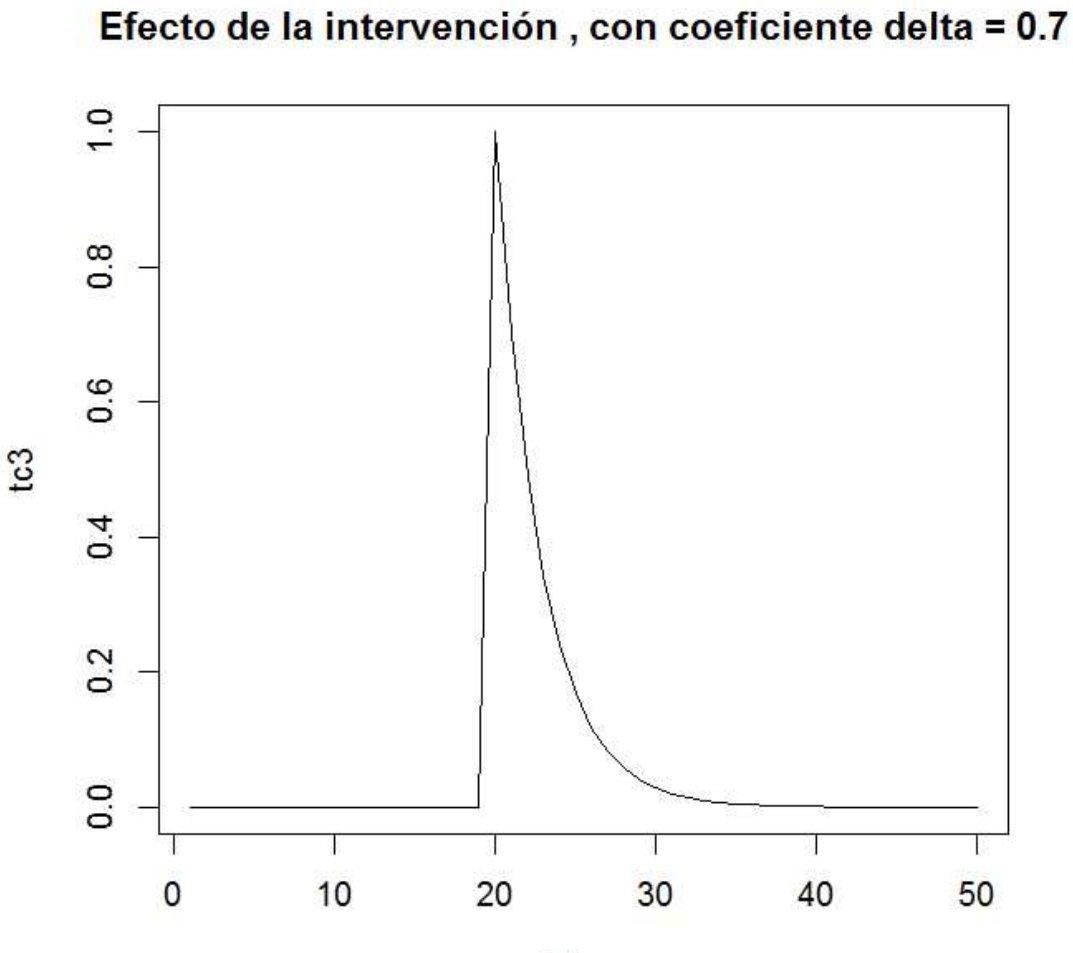
$$\begin{aligned} Z_h &= W_0 + y_h \\ Z_{h+1} &= W_1 + y_{h+1} \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_m &= W_m + y_{m+h} \end{aligned}$$

Variables Impulso. Función de repuesta a impulsos III

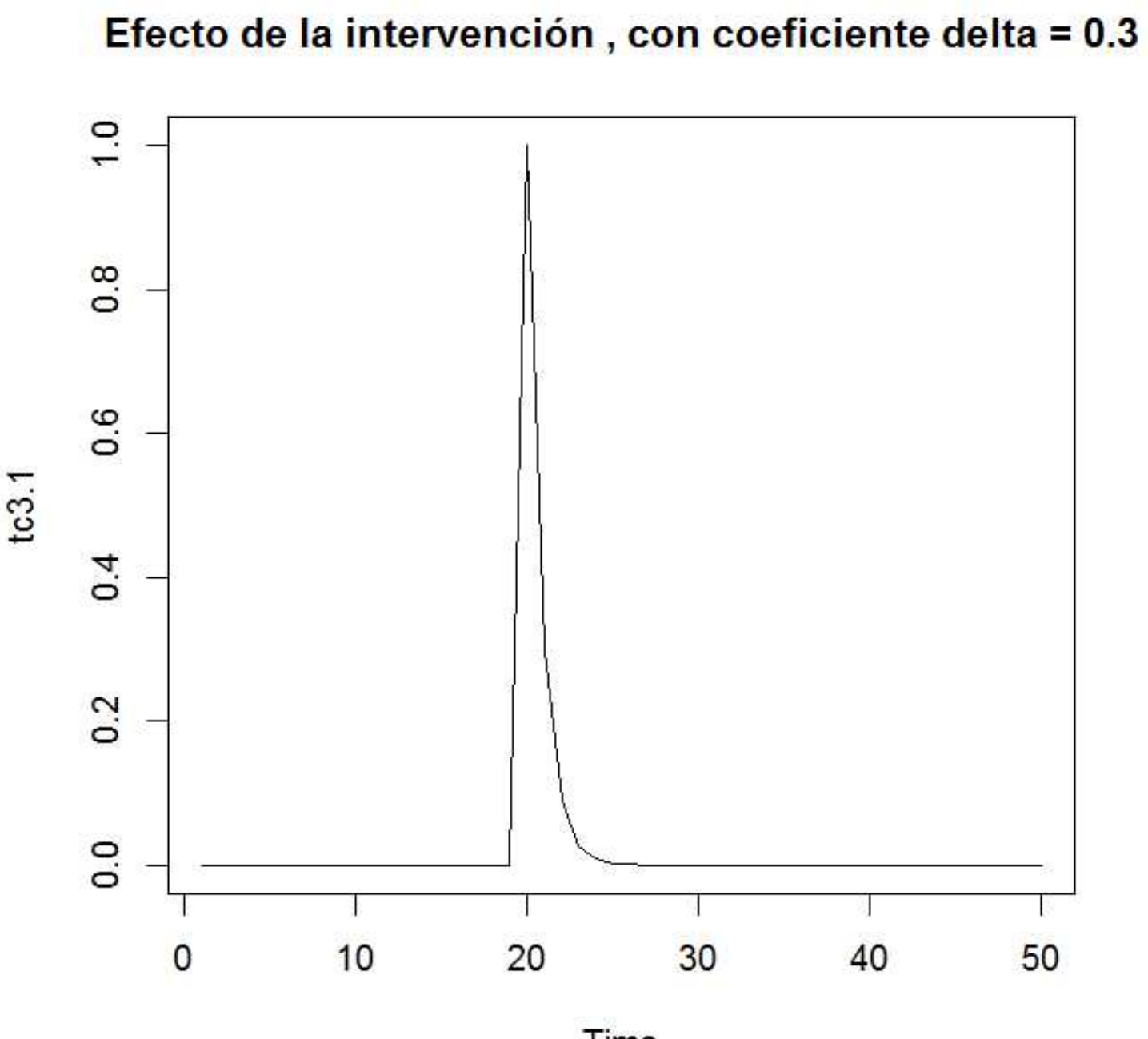
Un Modelo más general sería:

$$\Rightarrow \; Z_t = w(L)I_t^h + \Psi(L)\varepsilon_t \tag{6}$$

Con $w(L) = (W_0 + W_1 L + W_2 L^2 + ... + W_m L^m)$ que es el efecto dinámico que ha tenido el impulso, el que se representa mediante un polinomio de orden m, lo que permite extender el efecto sobre los m períodos siguientes a la intervención.



Variables Impulso. Función de repuesta a impulsos IV



Variables Impulso. Función de repuesta a impulsos V

Si el número de períodos afectados por el impulso es largo, estos efectos se pueden representar:

$$Z_t = \frac{W_0}{(1-\delta.L)}.(I_t^h) + \Psi(L)\varepsilon_t \quad con \;\; 0 < \delta < 1 \tag{7}$$

La función de respuesta a impulso es:

$$\frac{W_0}{1-\delta L} = W_0(1+\delta L+\delta^2 L^2+...) =$$

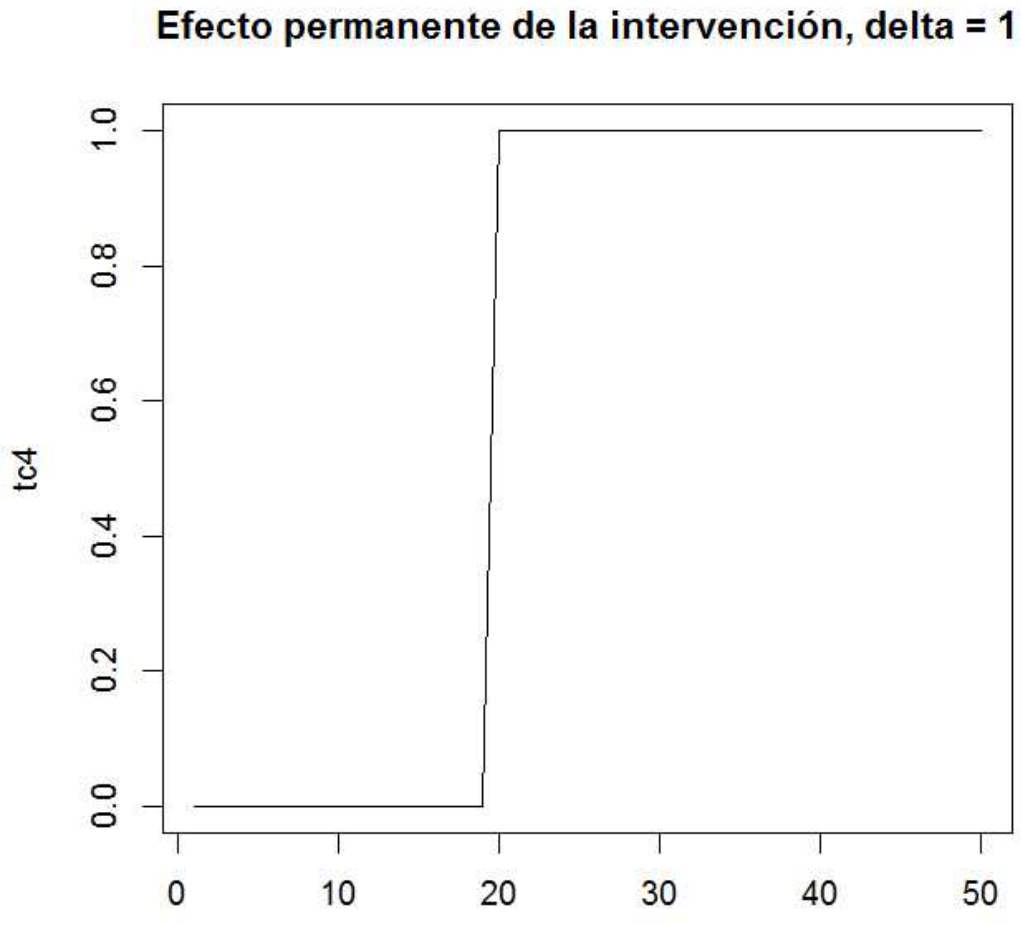
$$= (W_0 + W_1.L + W_2.L^2 + ... + W_m.L^m) \text{ con } W_k = W_0\delta^k \tag{8}$$

El efecto de la intervención se representa a través de un polinomio con pesos decrecientes, que tienden a cero a medida que nos alejamos del momento de la intervención.

Variables Escalón: Ganancia

Si se quiere modelar intervenciones que tienen un efecto permanente sobre la serie a partir de su ocurrencia, entonces las intervenciones se modelan con variables escalón.

$$S_t^h = \begin{cases} 0 & \text{t} < \text{h} \\ 1 & t \geq \text{h} \end{cases}$$



Variables Escalón: Ganancia I

El efecto de una variable escalón sobre la serie Y_t que sigue un modelo ARMA puede representarse mediante el modelo de intervención.

$$Z_t = w(L)S_t^h + \Psi(L)\varepsilon_t \tag{9}$$

Sea $w(L) = W_0$, un cambio de nivel de la serie a partir de un momento $t = h$, en ese caso todos los valores posteriores a $t = h$ están afectados por una cantidad constante W_0.

Llamaremos ganancia de un escalón a su efecto final en el largo plazo, es equivalente a la suma de los efectos parciales.

$$Ganancia = W(1) = W_0 + W_1 .... + W_m \tag{10}$$

La diferencia entre una variable impulso y una escalón es que el efecto de la primera es transitorio y el de la segunda es permanente.

No siempre se conoce a priori si el efecto del evento es permanente o transitorio, existen diversas alternativas para llegar a la mejor especificación, la metodología incluye tanto la detección como la clasificación, como por ejemplo el test propuestos por Chen y Liu (1993), como los métodos de saturación propuestos inicialmente por Hendry (1999), sesarrollados y ampliados posteriormente por Johansen, S. and B. Nielsen (2009),Doornik,J. (2009).

También es posible aplicar procesos más artesanales para la detección.

Efectos deterministas generales

Podemos incluir en el modelo un efecto determinista mediante una serie general X_t.Por ejemplo consideremos una serie de ventas mensuales que presumiblemente se puede ver afectada por la relación entre días laborables y feriados.

Estas variables indicadoras también podrán ser utilizadas para recoger otros efectos, como el efecto calendario. A modo de ejemplo, el número de días laborables en el mes es una variable importante a considerar a la hora de modelar series que representan actividad económica por ejemplo.

Es posible incluir como un regresor una variable que cuente los días laborables de cada mes.
El coeficiente de esta variable permitirá medir el efecto de un día laborable adicional en la variable a modelar.

Si los días de la semana implican diferentes niveles de actividad, es posible representar esta heterogeneidad, incluyendo siete variables explicativas que tengan en cuenta el número de lunes, el número de martes, etc. en ese mes.

También se puede sofisticar la representación y se puede diferenciar el número de lunes respecto a domingos, martes respecto a domingos, etc.

Efectos deterministas generales I

El programa TRAMO-SEATS genera variables determinísticas que se definen como el número de días laborables (lunes a viernes) menos el número de sábado y domingos multiplicado por $5/2$.

De esta manera la variable será cero si hay un número fijo de semanas completas en el mes, como un mes de 28 días sin feriados, ese mes tiene 20 días laborables y 8 no laborables $(20-8)*5/2$. La variable será positiva cuando haya relativamente más días laborables en el mes. La interpretación del coeficiente es el efecto de un día extra laborable sobre la relación 20-8. domingos multiplicado por $5/2$.

Similar estrategia se puede usar para recoger el efecto de los feriados móviles como turismo y carnaval, incluyendo una variable indicadora que cuenta la cantidad de días del feriado que caen en ese mes o trimestre.

La estimación de un modelo de intervención se realiza por *Máxima Verosimilitud*, la estimación incluye ahora los parámetros del modelo **ARIMA** y los de la intervención y puede interpretarse como una estimación en etapas.

$$Z_t = W_0 I_t^h + \frac{1}{(1-\phi L)}\varepsilon_t \tag{11}$$

1. Partiendo de una estimación inicial del efecto de la intervención W_0 se construye la serie corregida del efecto de la intervención:

$$y_t = Z_t - \hat{w}_0 I_t^h \tag{12}$$

2. Se estiman los parámetros **ARIMA** de la serie corregida Y_t maximizando la *Función de Verosimilitud*. En este ejemplo se estima el parámetro ϕ del $AR(1)$ $y_t = \phi y_{t-1} + \varepsilon_t$

1. Se calculan los residuos del modelo con los parámetros estimados y se utilizan para estimar el efecto de la intervención escribiendo los residuos como función de los efectos. Se multiplica toda la ecuación por la estructura del modelo **ARIMA** estimado.

$$\begin{aligned} Z_t &= W_0 I_t^h + \frac{1}{(1-\phi L)}\varepsilon_t \\ \left(1-\hat{\phi}L\right) Z_t - W_0 I_t^h &= \varepsilon_t \\ \hat{e}_t &= Z_t - \hat{\phi} Z_{t-1} \\ \hat{e}_t &= W_0 \left(1-\phi L\right) I_t^h + \varepsilon_t \end{aligned} \tag{13}$$

La última ecuación es la ecuación de regresión, el parámetro a estimar es el tamaño del efecto W_0
Sea $\xi_t = \left(1-\hat{\phi}L\right) I_t^h = I_t^h - \hat{\phi} I_{t-1}^h$, que es una variable que toma valor cero en todo su recorrido salvo en $\xi_h = 1$ y en $\xi_{h+1} = -\hat{\phi}$.

$$\hat{e}_t = W_0 \xi + \varepsilon_t \;\rightarrow\; \text{se estima } W_o; \hat{w}_0 \tag{14}$$

Con este valor se vuelve a iterar hasta obtener **convergencia**

Valores Atípicos

Con frecuencia ocurren en las series hechos puntuales que *desconocemos*. Las observaciones afectadas por esos eventos pueden representar una estructura diferente de las demás observaciones y aparecer como datos **atípicos**, no generados aparentemente como las demás observaciones.

Es importante poder identificar estas situaciones desconocidas, porque por ejemplo sus efectos impactan tanto en el proceso de identificación, estimación como en la predicción. Tienen el poder de *sesgar* la estimación de los parámetros, modificar la distribución de la serie y producir malas predicciones.

Si el suceso ha ocurrido sobre el final de la muestra y alguna de esas observaciones se utiliza para generar las predicciones, éstas serán malas.

Chang,Tiao y Chen (1988) proponen un procedimiento iterativo para la detección de los outliers y la estimación del modelo. Chen y Lui (1993) mejoran el procedimiento para estimar en forma conjunta los parámetros y los efectos de los outliers.

Atípicos aditivos (AO)

Diremos que ha ocurrido un atípico aditivo, **AO** sobre una serie temporal en el momento h si el valor de la serie se genera en ese instante siguiendo un proceso diferente que el resto.

El modelo que seguirá la serie Z_t observada si ha sido afectada por un AO en t es :

$$Z_t = \begin{cases} y_t & t \neq h \\ y_t + W_A & t = h \end{cases}$$

donde la serie Y_t sigue un modelo ARIMA.Por lo que el modelo que sigue la serie observada es:

$$Z_t = W_A I_t^h + \Psi(L)\varepsilon_t \tag{15}$$

con

$$I_t = \begin{cases} 0 & t \neq h \\ 1 & t = h \end{cases}$$

El nivel de la serie observada resulta afectada en $t = h$ en que se produce el atípico. El efecto no depende de la forma del **filtro ARIMA** Puedo escribir $\Pi(L)y_t = \varepsilon_t$ que es el efecto sobre la serie de las innovaciones

$$\Pi(L)(Z_t - W_A I_t^h) = \varepsilon_t \tag{16}$$

Atípicos aditivos (AO) I

A partir de $t = h$ en que ocurre el atípico la cadena de innovaciones se ve afectada en función de los coeficientes del filtro $\Pi(L)$.

La huella del AO en el nivel de la serie es la alteración del valor en el punto.Cuando desconozcamos su presencia y pretendemos construir un modelo ARIMA, su presencia puede notarse en los residuos del modelo.

Efectos en los Residuos

Vamos a suponer que conocemos los verdaderos valores de los parámetros del proceso y para simplificar supondremos que el verdadero proceso es un AR(1).Ignoramos la existencia del atípico en h y se estiman los residuos suponiendo que todas las observaciones son generadas por el mismo proceso.

$$Y_t = \Phi Y_{t-1} + \varepsilon_t \text{ tal que } Y_t \sim AR(1) \tag{17}$$

$$\begin{aligned} e_t &= y_t - \phi y_{t-1} \\ (1-\phi L)\, y_{t-1} &= (1-\phi L)\, W_A I_t^h + \varepsilon_t \end{aligned} \tag{18}$$

Atípicos aditivos (AO) II

La relación entre los residuos calculados y las verdaderas innovaciones es

$$e_t = W_A I_t^h - \Phi W_A I_{t-1}^h + \varepsilon_t \tag{19}$$

Antes de que ocurra el **atípico** en $t = h$,$e_t = \varepsilon_t$ pero a partir de $t = h$

$$\begin{aligned} e_h &= W_A + \varepsilon_h \\ e_{h+1} &= -\Phi W_A + \varepsilon_{h+1} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ e_{h+j} &= \varepsilon_{h+j} \; j \geq 2 \end{aligned} \tag{20}$$

Un $AR(1)$ afectado por un AO tendrá solo 2 residuos afectados y el efecto en la segunda observación posterior al primer efecto del atípico es de signo contrario a la primera y con una magnitud que es el efecto inicial multiplicado por el parámetro del modelo.

Atípico Cambio de Nivel (LS)

Un efecto muy importante que puede ocurrir en una serie es un cambio de nivel, cuando la serie ha sufrido un cambio de nivel en h se puede escribir como:

$$\begin{aligned} Z_t &= W_L S_t^h + \Psi(L)\varepsilon_t \\ Z_t &= \frac{1}{(1-L)} W_L I_t^h + \Psi(L)\varepsilon_t \end{aligned} \tag{21}$$

donde S_t^h es una variable escalón del tipo

$$S_t = \begin{cases} 1 & t \geq h \\ 0 & \text{otro caso} \end{cases} \qquad Z_t = \begin{cases} y_t & t < h \\ y_t + W_L & t \geq h \end{cases}$$

Si el proceso es estacionario un cambio de nivel convertiría la serie en no estacionaria, ya que la esperanza de cada observación será μ antes del cambio y $\mu + W_L$ después.

Efectos en los Residuos

$$\begin{aligned} Z_t - W_L S_t^h &= \Psi(L)\varepsilon_t \\ \Pi(L)(Z_t - W_L S_t^h) &= \varepsilon_t \end{aligned} \tag{22}$$

La relación entre los residuos calculados con los verdaderos parámetros y las innovaciones es :

$$e_t = W_L \Pi(L) S_t^h + \varepsilon_t \tag{23}$$

$$e_t = \begin{cases} a_t & t < h \\ a_t + W_L I_j & t = h + j \end{cases}$$

con $\mathscr{L}_j = 1 - \Pi_1 - \Pi_2 ..... - \Pi_j$ coeficientes de $\mathscr{L}(L)\frac{\Pi(L)}{(1-L)}$ Concluimos que todos los residuos después del cambio están afectados pero el efecto depende :

- del modelo, el efecto será mayor para los modelos estacionarios que para los no estacionarios
- de la distancia entre el momento de ocurrencia h y el final

Atípico Cambio Transitorio (TC)

Un *atípico transitorio* es un suceso cuyo efecto en la serie observada perdura pero no permanece

$$Z_t = \frac{W_{TC}}{(1-\delta L)} I_t^h + \Psi(L)\varepsilon_t \tag{24}$$

- Si $\delta = 1 \Rightarrow$ Modelo para el cambio de nivel (LS)
- Si $\delta = 0 \Rightarrow$ Modelo para el cambio AO

El **efecto rampa** es el que sigue

$$Z_t = W_R R_t^h + \Psi(L)\varepsilon_t \tag{25}$$

$$R_t^h = \begin{cases} 0 & t < h \\ t+1-h & t \geq h \end{cases} \tag{26}$$

que introduce una **tendencia determinística** con pendiente W_R en la serie a partir de h.
Luego del impacto inicial, el efecto del atípico cae gradualmente. El parámetro δ determina la velocidad con la que el efecto del atípico desaparece en el nivel de la serie.

Los efectos de los *outliers* en el proceso de identificación, en especial en las FAC y FACP estimadas se estudian en Chang (1982) y se retoman en Chan (1995). Chang concluye que la existencia de atípicos puede causar sesgos sustanciales en las FAC y FACP estimadas , las que pueden inducir a un proceso de identificación con errores.
Outliers aditivos de magnitud importante pueden destruir la información de la FAC estimada.

Este sesgo depende del número, el tipo,la magnitud y la posición relativa de los atípicos.

Para muestras de tamaño moderado, el sesgo puede causar sub o sobre especificación del modelo si se usa el FAC y FACP estimadas para la identificación, sin tener en cuenta la presencia de los atípicos.

Los efectos de los outliers no solo afectan el proceso de identificación y estimación de los parámetros del modelo sino que, de acuerdo a su ubicación, las proyecciones de la serie pueden verse dramáticamente afectadas.

Si el atípico se ubica en la mitad de la muestra, su efecto en lo sustancial se acota al proceso de identificación y estimación.
Pero en los casos de los cambios de nivel, donde puede llegar a confundirse un proceso estacionario, con cambio de nivel con un proceso integrado (con raíz unitaria).

La construcción de las predicciones se ve alterada por la vía del error en el modelo estimado y sus consecuencias en la mala estimación de la dinámica futura de la serie.

Si el atípico se da sobre el final de la muestra, se pueden distinguir dos casos:

1. Si el atípico se da sobre el final de la muestra, se suma la incertidumbre respecto a cuál puede ser la duración del efecto del mismo y por tanto compromete su clasificación y su correcta modelización a futuro.
2. Si el atípico se da próximo al final de la muestra, hay más elementos para poder clasificarlo, pero los parámetros del modelo estimado se afectarán.

En Trivez (1994) se analiza con detalle estos efectos.

Mecanismos de identificación de los atípicos

En la práctica la posición y la naturaleza de los atípicos es en general desconocida y se requiere un procedimiento de identificación y clasificación para poder estimar sus efectos. Para cada atípico, se requiere:

- Detectar el momento en que se registra el atípico.
- Clasificar el tipo de atípico por la duración de su efecto (transitorio o permantente. AI,IO,TC,LS).
- Estimar la magnitud del efecto sobre la serie.

Hay diversos procedimientos diseñados para encontrar atípicos de tipo aditivo, innovativo, cambio transitorio y cambio de nivel.

En Chen y Liu (1993) se presenta una procedimiento para la estimación conjunta de los parámetros del modelo y los efectos de los outliers.
Este procedimiento consiste en identificar la localización y analizar la modelización del efecto, su intensidad y duración.
En un proceso sucesivo de identificación (especificación tentativa de la forma de modelización),estimación del modelo y diagnóstico del mismo.

Mecanismos de identificación de los atípicos I

Chen y Liu advierten que permanecen algunos puntos por resolver como:

- La presencia de atípicos pueden resultar en modelos inapropiados
- Incluso si el modelo se especifica adecuadamente la presencia de atípicos pueden producir sesgo en la en las estimaciones de los parámetros y afectar la eficiencia de la detección del atípico. Una dificultad que plantea el procedimiento de Box y Tiao (1975) es que tanto la clasificación como la localización de los outliers puede cambiar en las diferentes iteraciones.
- Algunos outliers pueden no identificarse debido a un efecto enmascaramiento.

Método de saturación

Las intervenciones de política así como otros eventos que generen cambios transitorios o permanentes en la trayectoria de la serie pueden generar cambios en la función de distribución, alterando la dependencia temporal de los datos y causando fallos sistemáticos en las predicciones. Es bien sabido que la ocurrencia de eventos extraordinarios (outliers) puede dar lugar a distribuciones asimétricas y con colas pesadas, aunque el proceso generador de datos no tenga una distribución de ese tipo.

Entre la familia de procedimientos disponibles Hendry,D. propone una metodología que se denomina estimación por saturación.

Estas metodologías resuelven lo que aparecía como un problema sin solución, relacionado con el proceso de selección de las variables a retener en el modelo, partiendo de una realidad en que pueden existir más variables que observaciones.

La primera solución para este problema fue propuesta por Hendry (1999), basado en una saturación por ventanas temporales. Con el paso del tiempo, el método de saturación amplió su base. Desde los trabajos pioneros, se desarrollaron el método IIS (impulse indicator saturation) , el SIS (step indicator saturation), el TIS (trend indicator saturation) hasta el método MIS (Multiplicative indicator saturation).

Método de saturación I

En el caso del **IIS** se crea un conjunto completo de variables indicatrices, una para cada observación de la muestra.

EL **SIS** considera variables de tipo escalón (cambio permanente de nivel) para cada observación, generando un índice acumulado de variables tipo impulso.

A través del **TIS**, se pueden capturar quiebres en la tendencia, saturando el modelo mediante indicadores de tendencia.

Este sistema es parte de un algoritmo de selección de modelos que se denomina **Autometrics** que incluye un procedimiento que permite realizar las estimaciones en casos donde hay más variables que observaciones.

La idea que hay detrás de esta metodología es que los quiebres pueden darse en cada punto del tiempo, aunque en el proceso de modelización se retendrán, solamente, aquellos que resultan estadísticamente significativos.

Bibliografía utilizada para elaborar estas presentaciones

- Box,G and Tiao, G (1975)- «Interventiona Analysisi with Applications to Economic and Enviromental Problems».*Journal of the American Statistical Association*. Vol.70, N° 349. Theory and Method Section.
- Chen, C. y Liu, L., (1993)- "Joint Estimation of Model Parameters and Outlier Effects in Time Series", *Journal of the American Statistical Association*, 88, 284-297.
- Doornik,J. (2009) Autometrics . The Methodology and Practice of Econometrics. Edited by Jenifer Castle y Neil Shepard. Oxford University Press
- Gómez,V. Taguas,D. (1995) - "Detección y corrección automática de outliers con TRAMO: Una aplicación al IPC de bienes industriales no energéticos", D-95006. Documento de Trabajo de la Dirección General de Planificación.Ministerio de Hacienda y Función Pública. Gobierno de España.

Bibliografía utilizada para elaborar estas presentaciones

- Hendry, D. F. (1999) An econometric analysis of US food expenditure, 1931-1989.
- Johansen, S. and B. Nielsen (2009). An analysis of the indicator saturation estimator as a robust regression estimator. The Methodology and Practice of Econometrics. Edited by Jenifer Castle y Neil Shepard. Oxford University Press.
- Peña, D. (2005) «*Análisis de series temporales*» Alianza Editorial.
- Trivez,J.(1994) «Efectos de los distintos tipos de outliers en las predicciones de los modelos ARIMA». *Estadística Española. Vol. 36 N° 135*
- Tsay,R.S. (1986) «Time Series Model Specification in the Presence of Outliers».*Journal of the American Statistical Association. Vol.81.N°393. Theory and Methods*