Maestría en Ciencias de Datos: Estadística

Análisis de Series de Tiempo

25/5/2020

Análisis de Series de Tiempo

El conjunto de técnicas de estudio de series de observaciones dependientes ordenadas en el tiempo se denomina Análisis de Series Temporales.

El instrumento de análisis que se suele utilizar es un modelo que permita reproducir el comportamiento de la variable de interés.

Los Modelos de Series Temporales pueden ser:

- Univariantes: sólo se analiza una serie temporal en función de su propio pasado.
- Multivariantes: se analizan varias series temporales a la vez.

Una serie temporal univariante consiste en un conjunto de observaciones de una variable Y. Si hay T observaciones, se denota por $Y_t, t \in \mathbb{Z}; t = 1, \ldots, T$.

El subíndice t indica el tiempo en que se observa el dato Y_t. Los datos u observaciones se suelen recoger a intervalos iguales de tiempo, es decir, equidistantes los unos de los otros; es el caso de series mensuales, trimestrales, etc.

Si las observaciones se recogen sólo en momentos determinados de tiempo, generalmente a intervalos iguales, nos referimos a una serie temporal discreta.

Si los datos se generan de forma continua y se observan de forma continua, por ejemplo, la temperatura, que se observa de forma continua en el tiempo por medio de aparatos físicos. En este caso denotamos la serie temporal por $Y_t, t \in \mathbb{R}$.

y se cuenta con un número infinito de observaciones. En este caso nos referiremos a una serie temporal continua.

Sin embargo, la mayoría de las series disponibles, se observan en tiempo discreto a intervalos iguales, aunque se puedan suponer generadas por algún proceso en tiempo continuo.

Cada uno de los datos, Y_t, puede representar una acumulación sobre un intervalo de tiempo de alguna cantidad subyacente, por ejemplo, lluvia diaria, o un valor tomado en un momento dado.

Los objetivos que se pueden plantear al analizar una serie temporal son muy diversos, pero entre los más importantes se encuentran:

a) Describir las características de la serie, en términos de sus componentes de interés. Por ejemplo, la tendencia de la serie o el comportamiento estacional.

b) Predecir futuros valores de las variables. Un modelo de series temporales univariante se formula únicamente en términos de los valores pasados de Y_t, y/o de su posición con respecto al tiempo (ningún modelo univariante puede ser tomado seriamente como un mecanismo que describe la manera en que la serie es generada: no es un proceso generador de datos).
Las predicciones obtenidas a partir de un modelo univariante no son más que extrapolaciones de los datos observados hasta el momento T. Sin embargo, son en muchas ocasiones son efectivas y proporcionan un punto de referencia con el que comparar el funcionamiento de otros modelos.

Si las observaciones sucesivas son dependientes, los valores futuros pueden ser predichos a partir de las observaciones pasadas.

Si una serie temporal se puede predecir exactamente, entonces es una serie determinista. La mayoría de las series son estocásticas porque el futuro sólo se puede determinar parcialmente por sus valores pasados, por lo que las predicciones exactas son imposibles y deben ser reemplazadas por la idea de que los valores futuros tienen una distribución de probabilidad que está condicionada al conocimiento de los valores pasados.

Hasta ahora se han utilizado métodos deterministas para analizar las series de tiempo. A continuación, se presentarán métodos bajo el enfoque estocástico, son métodos un poco más complejos y requieren de series de mayor tamaño.

Modelo ARIMA (autoregressive integrated moving average). Son modelos paramétricos que tratan de obtener la representación de la serie en términos de la interrelación temporal de sus elementos.

El modelo ARIMA necesita identificar los coeficientes y número de regresiones que se utilizarán. Este modelo es muy sensible a la precisión con que se determinen sus coeficientes.

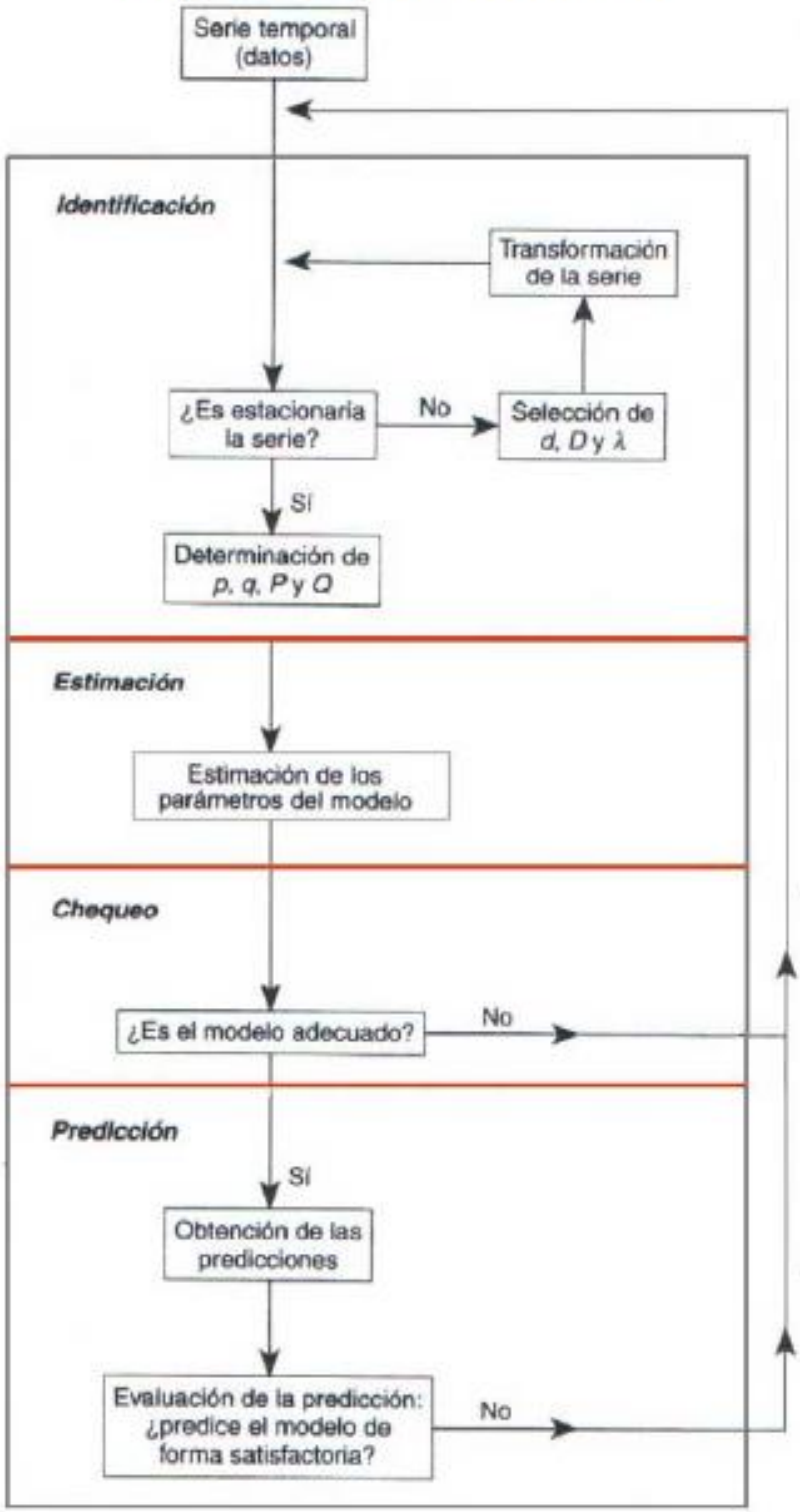
Se suele expresar como ARIMA(p,d,q) donde los parámetros p, d y q son números enteros no negativos que indican el orden de las distintas componentes del modelo, respectivamente las componentes autorregresiva, integrada y de media móvil.

El modelo **ARIMA** permite describir un valor como una función lineal de datos anteriores y errores debidos al azar, además, puede incluir un componente cíclico o estacional. Es decir, debe contener todos los elementos necesarios para describir el fenómeno. Box y Jenkins recomiendan como mínimo 50 observaciones en la serie temporal.

La metodología de Box y Jenkins se resume en cuatro fases:

- La ***primera fase*** consiste en identificar el posible modelo **ARIMA** que sigue la serie, lo que requiere:
  - Decidir qué transformaciones aplicar para convertir la serie observada en una serie estacionaria.
  - Determinar un modelo **ARMA** para la serie estacionaria, es decir, los órdenes p y q de su estructura autorregresiva y de media móvil.
- La ***segunda fase***: Seleccionado provisionalmente un modelo para la serie estacionaria, se pasa a la segunda etapa de estimación, donde los parámetros AR y MA del modelo se estiman por máxima verosimilitud y se obtienen sus errores estándar y los residuos del modelo.
- La ***tercera fase*** es el diagnostico, donde se comprueba que los residuos no tienen estructura de dependencia y siguen un proceso de ruido blanco. Si los residuos muestran estructura se modifica el modelo para incorporarla y se repiten las etapas anteriores hasta obtener un modelo adecuado.
- La ***cuarta fase*** es la predicción, una vez que se ha obtenido un modelo adecuado se realizan predicciones con el mismo.

METODOLOGÍA BOX-JENKINS



Pasos por seguir para el análisis de los datos:

1. Recogida de los datos.
2. Representación gráfica.
3. Transformación de la serie.
4. Eliminación de la tendencia.
5. Identificación del modelo.
6. Estimación de los coeficientes del modelo.
7. Análisis de los errores.
8. Selección del modelo.
9. Predicción.

Para hacer una representación gráfica, se usa un correlograma, que es una herramienta comúnmente usada para el control de aleatoriedad en un conjunto de datos. Esta aleatoriedad se determina calculando autocorrelaciones para los valores de datos en diferentes lapsos.

El instrumento fundamental para analizar las propiedades de una serie temporal en términos de la interrelación temporal de sus observaciones es el denominado coeficiente de autocorrelación que mide el grado de asociación lineal que existe entre observaciones separadas k períodos. Los coeficientes de autocorrelación proporcionan información sobre cómo están relacionadas entre sí las distintas observaciones de una serie temporal, lo que ayuda a construir el modelo apropiado para los datos.

El coeficiente de correlación entre dos variables x_1 y y_t se define como:

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{V(x)V(y)}}$$

Este coeficiente es adimensional y toma valores $-1 \leq \rho \leq 1$.
Si $\rho = 0$, no existe una relación lineal entre x y y

La autocorrelación muestral de una serie no estacionaria suele decrecer muy lentamente, mientas que para una serie estacionaria presenta el mismo tipo de comportamiento periódico que la serie en estudio.

3. Transformación de la serie:

- Estabilización de la varianza: Para estabilizar la variabilidad se suelen tomar logaritmos. Esta transformación funcionará bien cuando la variabilidad sea aproximadamente proporcional al nivel de la serie.

4. Eliminación de tendencia: Una forma sencilla de eliminar una tendencia aproximadamente lineal es diferenciar la serie, es decir, considerar la serie de diferencias entre una observación y la anterior en lugar de la serie original. Si x_t es una serie contenida en x, para calcular $\nabla x_t = x_t - x_{t-1}$

RUIDO BLANCO

Es un proceso puramente aleatorio, se define por las condiciones:

$$u = E(X_t) = 0\ ,\ g_0^2 = \text{var}(X_t)\ ,\ g_k = \text{cov}(X_t\ , X_{t+k}) = 0 \quad k = \cdots, -2, -1, 0, 1, 2, \cdots$$

En este tipo de procesos puramente aleatorios el **correlograma** se reduce a un segmento de longitud unitaria sobre el eje de ordenadas.

5. Identificación del modelo

MODELOS AUTORREGRESIVOS AR(p)

Un modelo autorregresivo **AR** describe una clase particular de proceso en que las observaciones en un momento dado son predecibles a partir de las observaciones previas del proceso más un término de error. El caso más simple es el ARIMA(1,0,0) o AR(1) o de primer orden, cuya expresión matemática es:

$$AR(1) \equiv X_t = \phi_1 X_{t-1} + a_t$$

El proceso autorregresivo de orden p, representado por ARIMA(p,0,0) o simplemente por AR(p):

$$AR(p) \equiv X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t$$

que puede ponerse, mediante el operador de cambio retroactivo B, en la forma:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\ X_t = a_t \qquad B^k (X_t) = X_{t-k}$$

- ♣ Un proceso autorregresivo AR(p) es **estacionario** si las raíces del polinomio en B dado por: $(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ caen fuera del círculo unidad. Esta condición es equivalente a que las raíces de la ecuación: $x^p - \phi_1 x^{p-1} - \phi_2 x^{p-2} - \cdots - \phi_{p-1} x - \phi_p = 0$ sean todas inferiores a uno en módulo.

- ♣ Un proceso autorregresivo siempre es **invertible**.

MODELO DE MEDIAS MÓVILES Ma(q)

Un modelo de medias móviles **MA** describe una serie temporal estacionaria. En este modelo el valor actual puede predecirse a partir de la componente aleatoria de este momento y, en menor medida, de los impulsos aleatorias anteriores. El modelo ARIMA(0,0,1), también denotado por MA(1), viene dado por la expresión:

$$X_t = a_t - v_1 a_{t-1}$$

El proceso de medias móviles de orden q, representado por ARIMA(0,0,q) o también por Ma(q), viene dado por la expresión:

$$X_t = a_t - v_1 a_{t-1} - v_2 a_{t-2} - \cdots - v_q a_{t-q}$$

que puede ponerse, mediante el operador de cambio retroactivo B, en la forma:

$$X_t = (1 - v_1 B - v_2 B^2 - \cdots - v_q B^q)\, a_t$$

- ♣ Un proceso de medias móviles es siempre **estacionario**.
- ♣ Un proceso de medias móviles Ma(q) es **invertible** si las raíces del polinomio en B definido por $(1 - v_1 B - v_2 B^2 - \cdots - v_q B^q)$ caen fuera del círculo unidad. Esta condición es equivalente a que las raíces de la ecuación $x^q - \phi_1 x^{q-1} - \phi_2 x^{q-2} - \cdots - \phi_{q-1} x - \phi_q = 0$ sean todas inferiores a uno en módulo.

MODELOS ARMA (p, q)

Una extensión natural de los modelos **AR(p)** y **MA(q)** es un tipo de modelos que incluyen tanto términos autorregresivos como de medias móviles y se definen como **ARIMA(p, 0, q)**. Se representan por la ecuación:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t - \nu_1 a_{t-1} - \nu_2 a_{t-2} - \cdots - \nu_q a_{t-q}$$

que puede ponerse de la forma:

$$X_t - \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} = a_t - \nu_1 a_{t-1} - \nu_2 a_{t-2} - \cdots - \nu_q a_{t-q}$$

es decir, $X_t (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) = a_t (1 - \nu_1 B - \nu_2 B^2 - \cdots - \nu_q B^q)$

El proceso **ARMA(p, q)** es **estacionario** si lo es su componente autorregresiva, y es **invertible** si lo es su componente de medias móviles.

- Un modelo **ARMA(p, q)** es **estacionario** si las raíces del polinomio definido por $(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ caen fuera del circulo unidad. Esta condición es equivalente a que las raíces de la ecuación:

  $x^p - \phi_1 x^{p-1} - \phi_2 x^{p-2} - \cdots - \phi_{p-1} x - \phi_p = 0$ sean todas inferiores a uno en módulo.

- Un modelo **ARMA(p, q)** es **invertible** si las raíces del polinomio en B definido mediante $(1 - \nu_1 B - \nu_2 B^2 - \cdots - \nu_q B^q)$ caen fuera del circulo unidad. Esta condición es equivalente a que las raíces de la ecuación:

  $x^q - \phi_1 x^{q-1} - \phi_2 x^{q-2} - \cdots - \phi_{q-1} x - \phi_q = 0$ sean todas inferiores a uno en módulo.

MODELOS ARIMA (p, d, q)

Un modelo ARIMA(0, d, 0) es una serie temporal que se convierte en ruido blanco (proceso puramente aleatorio) después de ser diferenciada **d** veces.

El modelo (0, d, 0) se expresa mediante: $(1-B)^d X_t = a_t$

El modelo general ARIMA(p, d, q) denominado proceso autorregresivo integrado de medias móviles de orden p, d, q, toma la expresión:

$$(1-\phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1-B)^d \; X_t = (1-\nu_1 B - \nu_2 B^2 - \cdots - \nu_q B^q)\, a_t$$

Este modelo se puede aplicar a cualquier modelo. Si hay una componente p,d,q, igual a cero, se elimina el término correspondiente.

Los modelos cíclicos o estacionales son aquellos que se caracterizan por oscilaciones cíclicas, también denominadas variaciones estacionales. Las variaciones cíclicas a veces se superponen a una tendencia secular.

Las series con tendencia secular y variaciones cíclicas pueden representarse mediante los modelos **ARIMA(p, d, q)(P, D, Q)**. El primer paréntesis **(p, d, q)** se refiere a la **tendencia secular o parte regular** de la serie y el segundo paréntesis **(P, D, Q)** se refiere a las **variaciones estacionales**, o **parte cíclica** de la serie temporal.

$$\text{ARIMA}\ \underbrace{(p,d,q)}_{\text{Parte regular o tendencia}}\ \underbrace{(P,D,Q)}_{\text{parte estacional}}$$

p,P: # parámetros **AR**

d,D: # de diferenciaciones **I**

q,Q: # parámetros **MA**

Regulares o estacionales

Ejemplos:

MA(1) o ARIMA (0,0,1)(0,0,0)

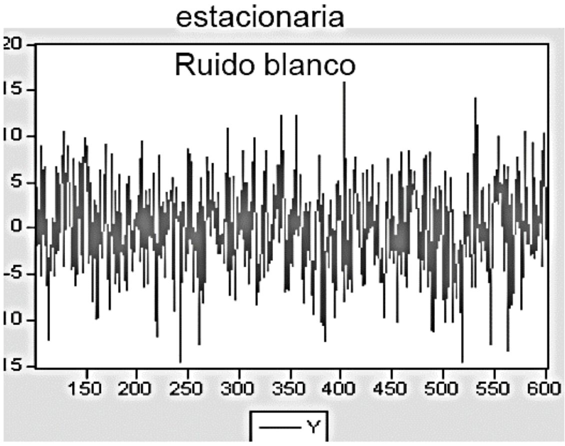
AR(1) o ARIMA (1,0,0)(0,0,0)

MA(1) MA(12) o ARIMA(0,0,1)(0,0,1)

Algunos casos especiales:

- **Ruido blanco** (white noise): $Z_t = a_t$ ARIMA(0,0,0)
  a_t es un componente aleatorio que tiene media cero y varianza constante σ^2_a. La historia no brinda ninguna información.
- **Camino aleatorio** (random walk): $Z_t = Z_{t-1} + a_t$ ARIMA (0,1,0)
  donde: $Z_t - Z_{t-1} = a_t$,con a_t ruido blanco
  el mejor pronóstico de t+1 es el valor observado en t.

estacionaria



no estacionaria