Series de tiempo

Diplomado en Ciencia de Datos

Dora Suárez

Serie de tiempo

Es una secuencia de valores, medidos en determinados momentos y ordenados de forma cronológica.

Las series de tiempo capturan información del mismo fenómeno en diferentes momentos de tiempo

Componentes de una serie de tiempo

Tendencia: Comportamiento general de la serie de tiempo

Estacionalidad: Comportamientos periódicos a corto plazo

Error: Comportamientos aleatorios

$$Y_t = \boxed{T_t + S_t} + \epsilon_t$$

Componente Determinística

Componente Aleatoria

Ejercicio

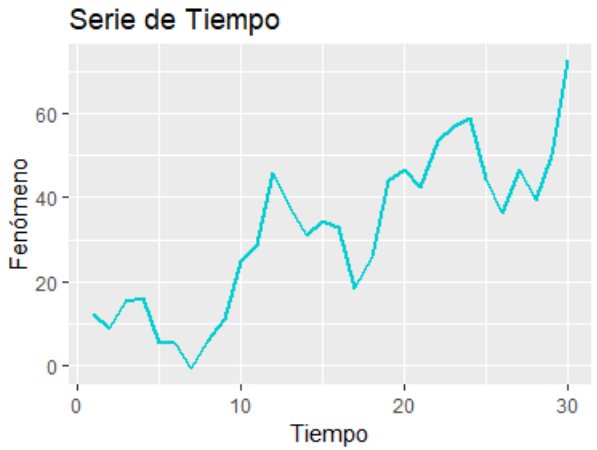
1. Crear en R una secuencia de 30 datos (almacenarlo en un vector x)
2. Crear un vector que contenga la recta t = 2x (almacenarlo en un vector t)
3. Crear un vector que contenga la siguiente sinusoidal: 12*sin(2*x/pi) (Almacenarlo en s)
4. Crear un vector de errores como: 4*rnorm(30) (almacenarlo en e)

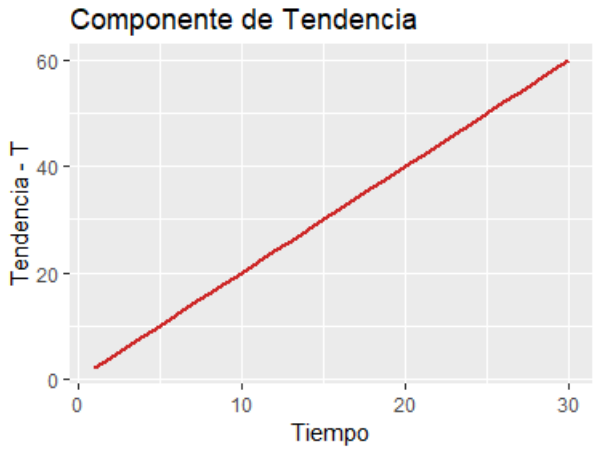
Ejercicio

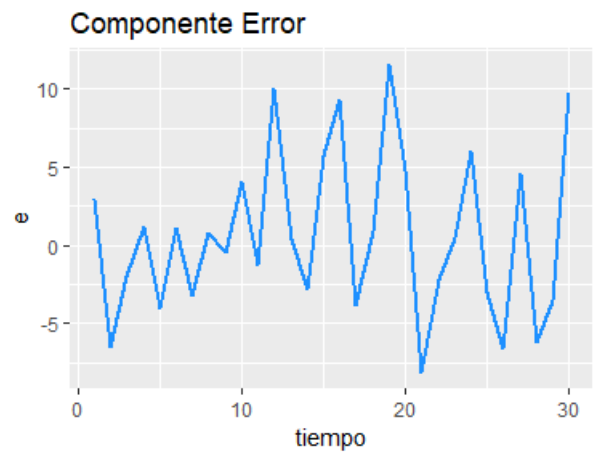
1. Calcular y = x + t + s + e

2. Hacer un gráfico de la tendencia, la estacionalidad, y la serie de tiempo completa (usar geom_line)

Componentes de una serie de tiempo







$$Y_t = T_t + S_t + \epsilon_t$$

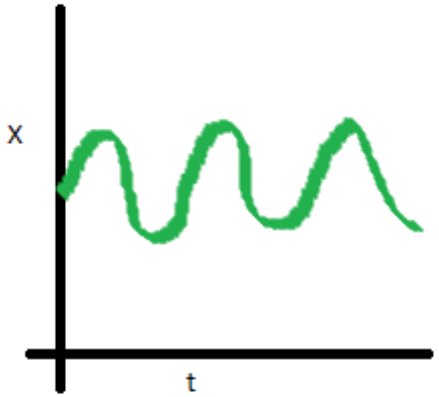
Series de tiempo estacionarias

1. El valor medio de las series de tiempo es constante en el tiempo, lo que implica que el componente de tendencia se anula.

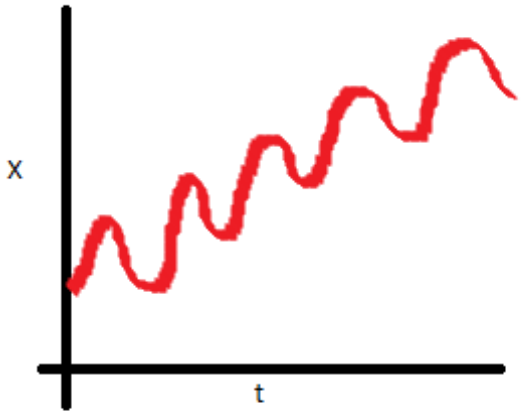
2. La varianza no aumenta con el tiempo.

3. El efecto de la estacionalidad es mínimo.

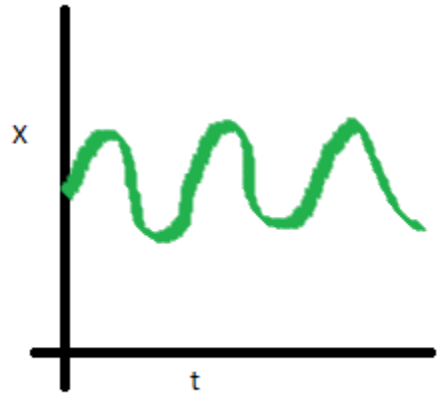
Series estacionarias y no estacionarias



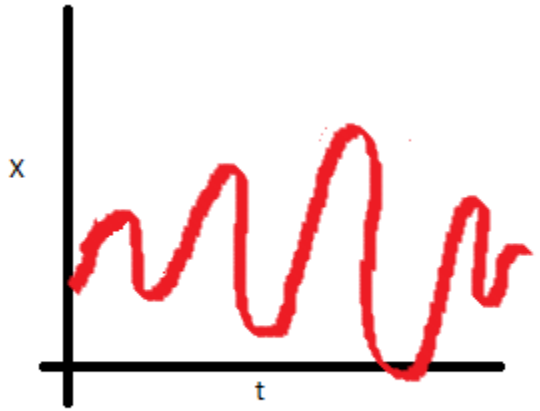
Stationary series



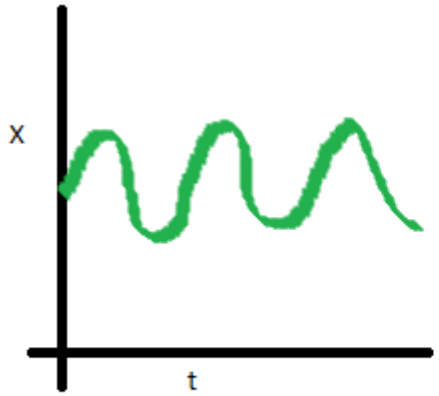
Non-Stationary series



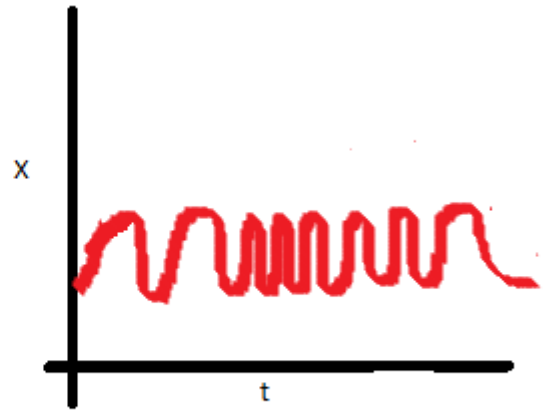
Stationary series



Non-Stationary series



Stationary series



Non-Stationary series

Media móvil

Es el resultado del promedio de conjuntos de observaciones de forma agrupada, sirve para suavizar líneas de tendencia

Ej: x = [1, 4, 9, 15, 22, 31]

Media móvil tamaño 2:

Y = [2.5, 6.5, 12, 18.5, 26.5]

Medias Móviles – Ejercicio

Ejecutar las siguientes líneas de Código ¿Qué observa?

```
x <- 1:20 %>% .^2 + 40*rnorm(20)
plot(x, type = "l")
y = rollmean(x, k = 3) # Calculo del promedio móvil
plot(y, type = "l")
```

Descripción de una serie de tiempo

1. Indicarle al software que la base de datos es de tipo temporal
2. Analizar los posibles ciclos existentes en la serie temporal
3. Descomponer la serie en cada una de las partes (Tendencia, estacionalidad y error)

Primer conjunto de datos

```
gas = scan('http://verso.mat.uam.es/~joser.berrendero/datos/gas6677.dat')
plot(gas)

gas.ts = ts(gas, start = c(1966,1), frequency = 12)
plot(gas.ts)

boxplot(gas.ts ~ cycle(gas.ts))

cycle(gas.ts)

gas.ts.desc = decompose(gas.ts)
plot(gas.ts.desc, xlab='Año')
```

Transformaciones de una serie de tiempo

1. Estabilización de la varianza -> Logaritmo

2. Eliminación de la tendencia -> Primera diferencia finita

$$diff(x_t) = x_t - x_{t-1}$$

3. Eliminar la estacionalidad -> Doceava diferencia finita

$$diff(x_t) = x_t - x_{t-12}$$

Ejercicio: Calcular las siguientes transformaciones ¿Qué observa?

```
plot(log(gas.ts))
x = log(gas.ts)
dif1.x = diff(x)
plot(dif1.x)
dif12.dif1.x = diff(dif1.x, lag=12)
plot(dif12.dif1.x)
```

Modelaje de la tendencia

$$Y_t = T_t + S_t + \epsilon_t$$

Lineal

$$T_t = \beta_0 + \beta_1 t$$

Cuadrático

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

Cubico

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$$

Exponencial

$$T_t = \exp(\beta_0 + \beta_1 t)$$

Logístico

$$T_t = \frac{\beta_2}{1 + \beta_1 \exp(-\beta_0 t)}$$

Como hacer que una serie sea estacionaria?

1. Aplicando logaritmos

2. Diferencias finitas (para corregir estacionariedad y tendencia)

Segunda base de datos

bit.ly/2Pqh8Sk

Ventas al menudeo en dólares Estadounidenses desde enero de 1955 de una empresa de confecciones

Pronósticos basados en la tendencia

$$\hat{Y}_{t+j} = E(Y_{t+j}|Y_1, Y_2, \dots, Y_t)$$

1. Dividir los datos en entrenamiento y prueba
2. Calcular los modelos correspondientes como regresiones lineales
3. Medir el desempeño del modelo con criterios como AIC y BIC
4. Seleccionar el modelo y probarlo con los datos de prueba

Suavizadores – Descomposición

Regresión Local Loess: Busca encontrar una línea de tendencia basada en métodos no paramétricos

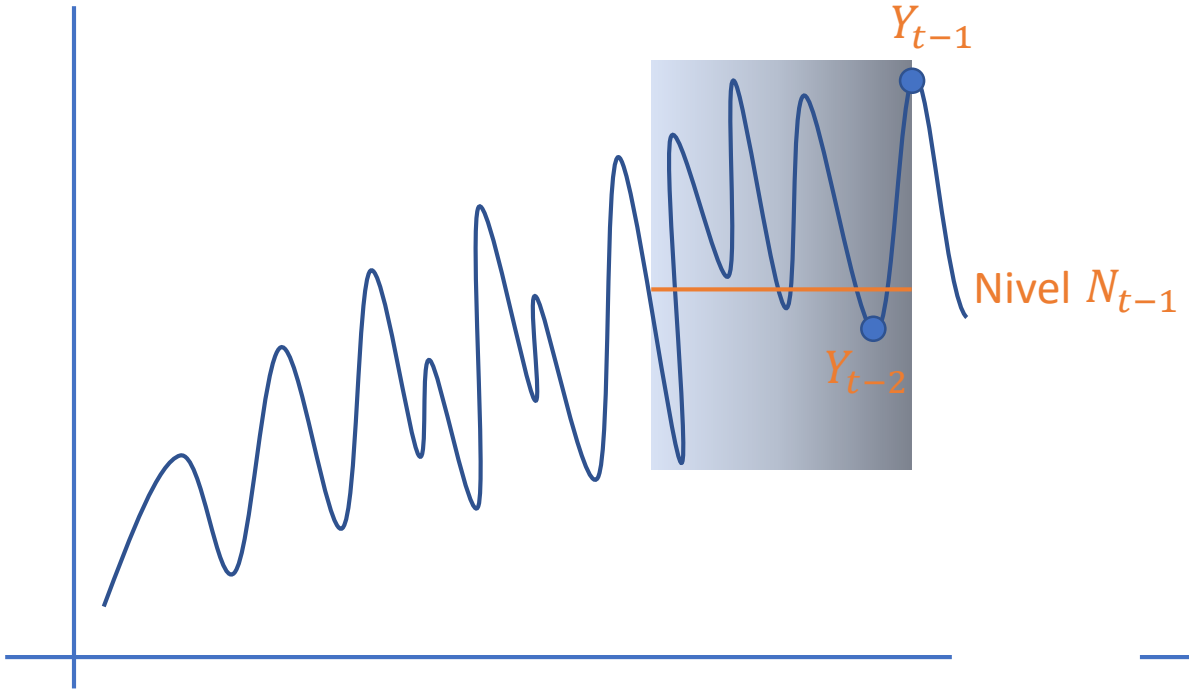
1. Se escoge $q \in \mathbb{Z}^+$ tal que $q \leq n$.
2. Se escogen los q valores x_i más cercanos a x
3. Defina $w(x) = \begin{cases} (1-x^3)^3 & , \ 0 \leq x \leq 1 \\ 0 & , \ x \geq 1 \end{cases}$
4. Defina $\lambda_q(x)$ la distancia de x al x_i más alejado entre los q escogidos.
5. Defina $v_i(x) = w\left(\frac{|x_i - x|}{\lambda_q(x)}\right)$, para cada x_i escogido.
6. Ajuste $Y_i = a + bx_i$ ó $Y_i = a + bx_i + cx_i^2$ con MCP ponderando cada x_i con $v_i(x)$.
7. Defina $g(x)$ como el valor estimado de la regresión en x.

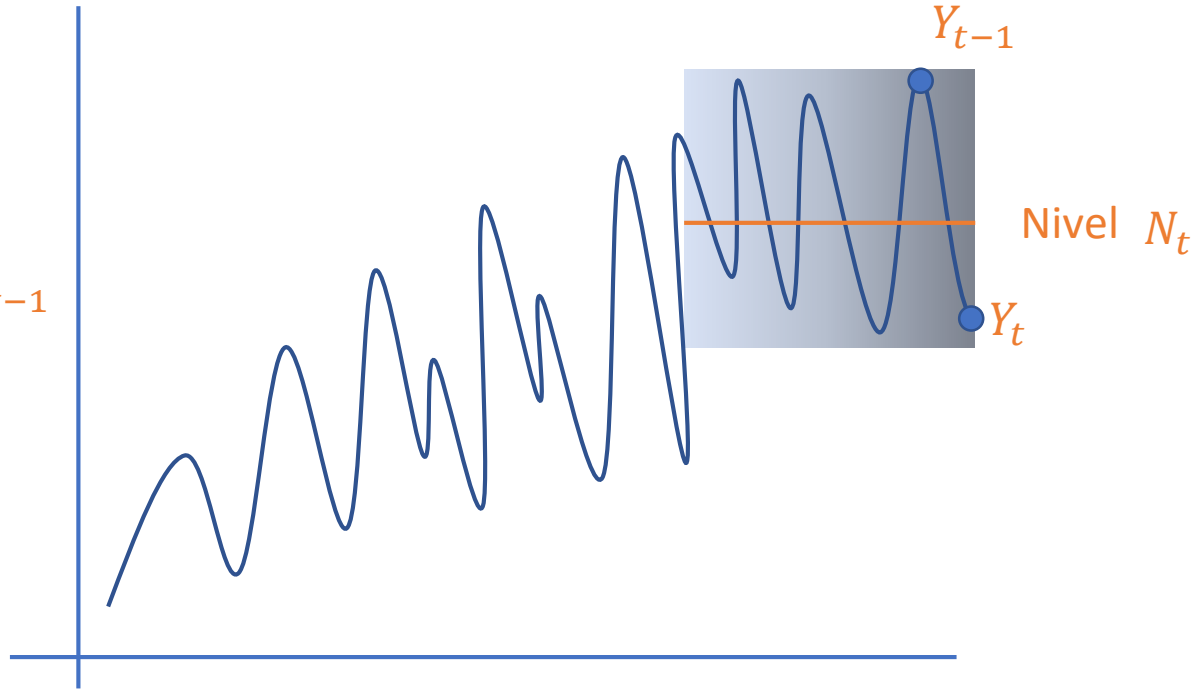
Algoritmos de suavizamiento exponencial

Idea: La serie tiene un nivel N_t desconocido que no necesariamente es constante pero las variaciones de el pueden ser imperceptibles

Objetivo: Estimar el Nivel

Algoritmos de suavizamiento exponencial





Algoritmos de suavizamiento exponencial

Se basan en formas de recurrencia:

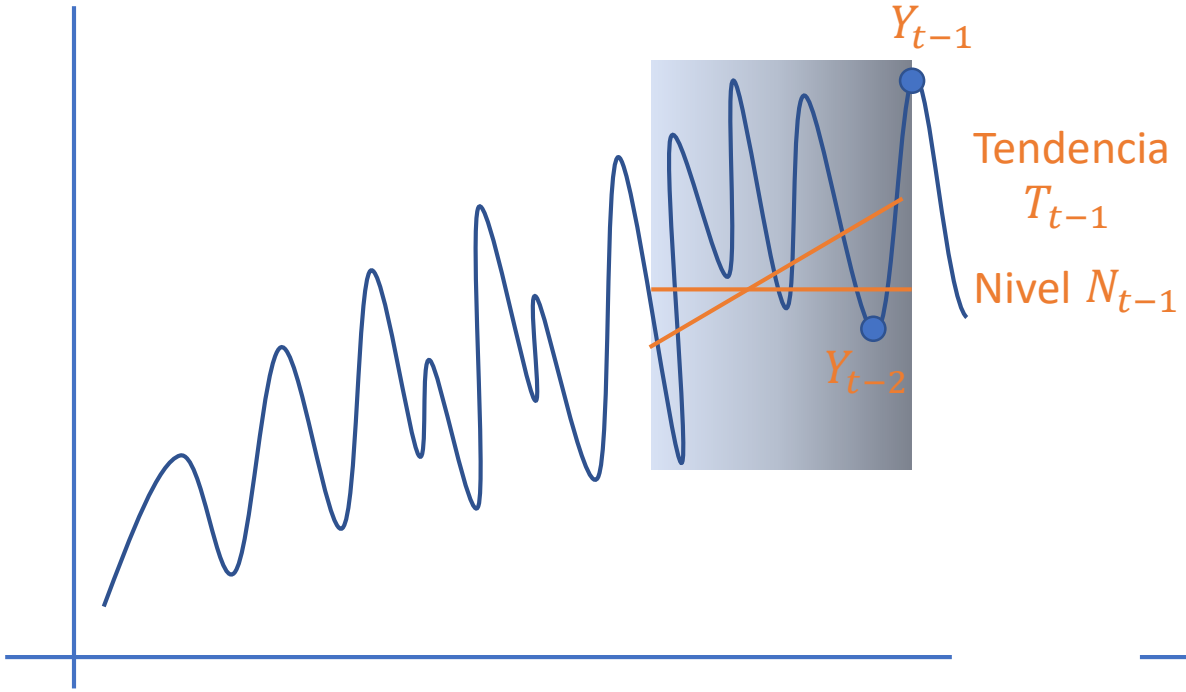
$$N_t = \alpha y_t + (1 - \alpha) N_{t-1}$$

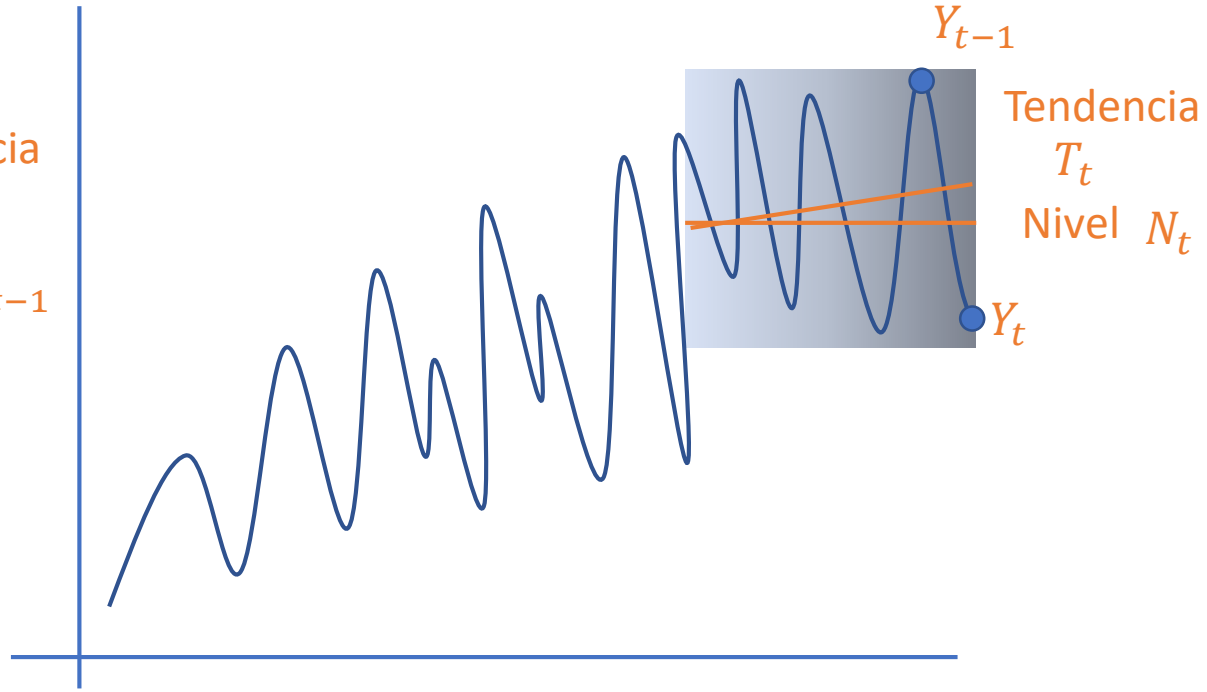
Se asume que la serie tiene un determinado nivel N_t asociado a la tendencia. Cuanto menor sea el valor de α, mayor peso es dado a la estimativa anterior. Suponemos que la mejor predicción que podemos tener de “mañana” es la observación que tenemos “hoy”

Ejercicio

1. Dividir la serie del precio de cierre de las acciones de Microsoft en entrenamiento y prueba, para ello dejar los últimos 20 días para hacer la predicción
2. Ajustar el algoritmo de suavizamiento exponencial para la serie de entrenamiento para 100 valores diferentes de Alpha
3. Para cada uno de los modelos ajustados, predecir el valor de la serie en los siguientes 20 días y calcular el error respecto a la serie real
4. Seleccionar el valor de Alpha que minimice el error

Algoritmo de Holt





Algoritmo de Holt

Se asume que la dinámica de la serie es determinada por dos componentes no observables que no necesitan ser fijas (nivel y tendencia)

Algoritmo:

$$N_t = \alpha y_t + (1-\alpha)(N_{t-1} + T_{t-1})$$
$$T_t = \beta(N_t - N_{t-1}) + (1-\beta)T_{t-1}$$

Inicialización: $N_2 = y_2; T_2 = y_2 - y_1$

Pronósticos: $\hat{y}_t(h) = N_t + hT_t, h = 1, 2, \ldots$

Ejercicio

Simular una serie de tiempo aleatoria con una tendencia (ver primer ejercicio)

```
x <- 1:30
t <- 2*x
s <- 12*sin(2*x/pi)
e <- 4*rnorm(30)
y <- t + s + e

datos <- data.frame(fenomeno = y,
                    tendencia = t,
                    estacionalidad = s,
                    error = e,
                    tiempo = x)
```

Ejercicio

Hacer una función que de forma recurrente calcule las predicciones para el algoritmo de Holt:

Inicialización:

N2 = y2

T2 = y2-y1

$$N_t = \alpha y_t + (1 - \alpha)(N_{t-1} + T_{t-1})$$
$$T_t = \beta(N_t - N_{t-1}) + (1 - \beta)T_{t-1}$$

Algoritmo de Holt

Ventajas:

Aprende de los errores

Predicciones sencillas

Desventajas:

No tiene en cuenta la estacionalidad

Las predicciones a largo plazo son malas

Algoritmo de Holt - Winters

Es una ampliación del modelo anterior, se tienen en cuenta las componentes estacionales

Algoritmo aditivo:

$$N_t = \alpha(y_t - F_{t-s}) + (1-\alpha)(N_{t-1} + T_{t-1})$$
$$T_t = \beta(N_t - N_{t-1}) + (1-\beta)T_{t-1}$$
$$F_t = \gamma(y_t - N_t) + (1-\gamma)F_{t-s}$$

Inicialización: $N_2 = y_2; T_2 = y_2 - y_1; F_1 = (y_1 - N_1); \dots F_s = (y_s - N_s)$

Pronósticos: $\hat{y}_t(h) = N_t + hT_t, +F_{t+h+s}\ h = 1, 2, \dots$

Algoritmo de Holt Winters

Ventajas:

Ajusta la tendencia a los errores de pronostico

Predicciones sencillas

Tiene en cuenta el efecto de la estacionalidad

Desventajas:

Las predicciones a largo plazo son malas

No tiene en cuenta cambios en las variaciones a lo largo del tiempo

Depende de la cantidad de observaciones para ajustar buenas predicciones de la estacionalidad

Algoritmo de Holt - Winters

Es una ampliación del modelo anterior, se tienen en cuenta las componentes estacionales

Algoritmo Multiplicativo:

$$N_t = \alpha(y_t)/F_{t-s} + (1-\alpha)(N_{t-1} + T_{t-1})$$
$$T_t = \beta(N_t - N_{t-1}) + (1-\beta)T_{t-1}$$
$$F_t = \gamma(y_t)/N_t + (1-\gamma)F_{t-s}$$

Inicialización: $N_2 = y_2; T_2 = y_2 - y_1; F_1 = (y_1 - N_1); \dots F_s = (y_s - N_s)$

Pronósticos: $\hat{y}_t(h) = N_t + hT_t, +F_{t+h+s}\ h = 1, 2, \dots$

Algoritmo de Holt Winters

Ventajas:

Ajusta la tendencia a los errores de pronostico

Predicciones sencillas

Tiene en cuenta el efecto de la estacionalidad

Tiene en cuenta las variaciones a lo largo del tiempo

Desventajas:

Las predicciones a largo plazo son malas

Depende de la cantidad de observaciones para ajustar buenas predicciones de la estacionalidad

Ejercicio

1. Consultar la base de datos co2 (precargada en R)
2. Describir la serie (separarla en su componente de tendencia, estacionalidad y error) ¿Qué observa?
3. Use el comando `stl(serie, s.window = "period")` y grafique los resultados
4. Que observa en el siguiente gráfico `monthplot(serie)`
5. Que observa en el siguiente gráfico `seasonplot(serie)`

Metodología de Box y Jenkins, 1970

1. Identificación y selección del modelo
   - ¿La serie es estacionaria? Si no es así, ¿qué debería hacer para que sea estacionaria?
   - ¿Cuántos periodos anteriores influyen en una observación de la serie?

Metodología de Box y Jenkins, 1970

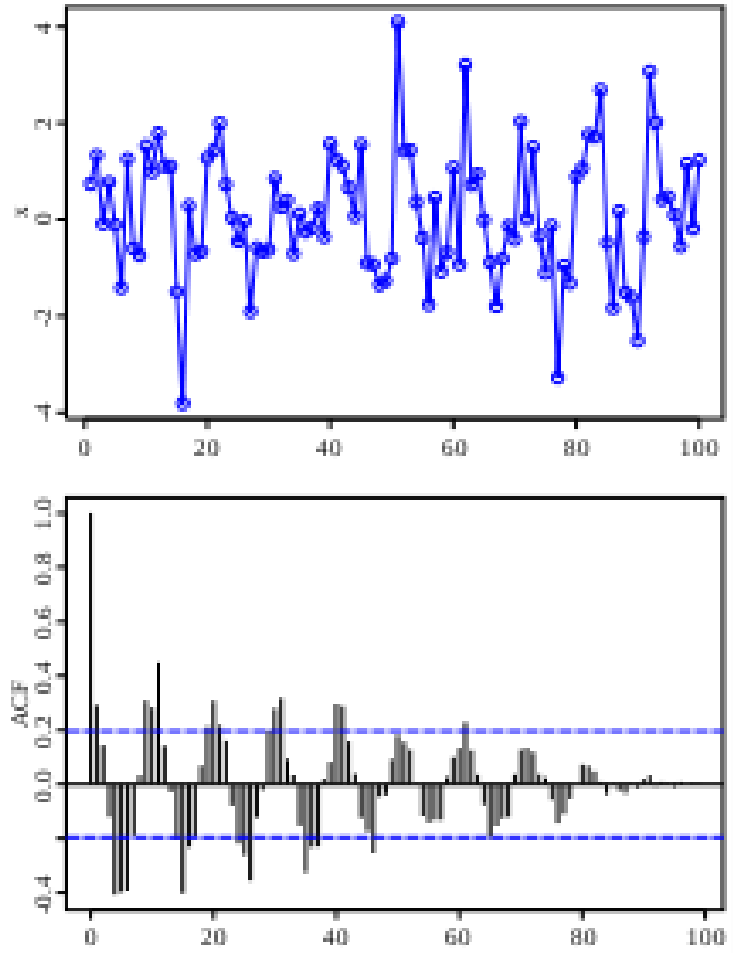
2. Estimación de parámetros
   - Máxima verosimilitud
   - Mínimos cuadrados no lineales

Metodología de Box y Jenkins, 1970

3. Comprobar el modelo
   - Ensayo con datos de test
   - ¿Los residuos son independientes?
   - ¿La media y varianza de los residuos son constantes en el tiempo?
   - ¿Los errores tienen comportamiento de ruido blanco estacionario?

Correlograma de una serie de tiempo

Una gráfica de autocorrelación está diseñada para mostrar si los elementos de una serie temporal están correlacionados positivamente, correlacionados negativamente o independientes entre sí.



Datos – Ejercicio

1- Consultar los datos ldeaths precargados en R ¿qué puede observar?

2- Haga una descripción de la serie de tiempo

3- Divida los datos en entrenamiento y prueba dejando los últimos 12 datos como los de prueba

Funciones de autocorrelación

La función de autocorrelación (ACF)

- Mide la correlación entre dos variables separadas por k periodos.

- Mide el grado de asociación lineal que existe entre dos variables del mismo proceso estocástico.

```
acf(serie_train)
```

Funciones de autocorrelación

La función de autocorrelación parcial (PACF)

- Mide la correlación entre dos variables separadas por k periodos cuando no se considera la dependencia creada por los retardos intermedios existentes entre ambas.

- Mide la autocorrelación que existe entre dos variables separadas k períodos descontando los posibles efectos debidos a variables intermedias

```
pacf(serie_train)
```

Ruido Blanco

Diremos que una serie $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ es un ruido blanco si:

[1] La media de la serie es cero:

$$E(\varepsilon_t) = 0$$

[2] La varianza de la serie es constante

$$V(\varepsilon_t) = \sigma^2$$

Ruido Blanco

Diremos que una serie $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ es un ruido blanco si:

[3] La serie es no auto-correlacionada

$$cov(\varepsilon_t, \varepsilon_{t+h}) = 0$$

[4] La serie tiene distribución aproximadamente normal

$$\varepsilon_t \underset{a}{\sim} N(\mu, \sigma^2)$$

Series de tiempo estacionarias en covarianza

La función de autocorrelación en una serie de ruido blanco es:

$$\rho(k) = \begin{cases} 1 & si \quad k = 0 \\ 0 & en\ otro\ caso \end{cases}$$

Donde $\rho(k)$ es la función de autocorrelación de la serie de tiempo entre el tiempo t y el tiempo $t + k$

Bandas de Barlett

En una serie estacionaria, se espera que con aproximadamente una probabilidad del 95% la correlación este entre:

$$\left[-\frac{2}{\sqrt{n}}, \frac{2}{\sqrt{n}}\right]$$

Procesos de medias móviles de media q

El valor actual depende de forma lineal de perturbaciones anteriores de la serie de tiempo

Modelo medias móviles de orden 1

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$
$$\varepsilon_t \sim RB(0, \sigma^2)$$

Modelo medias móviles de orden q

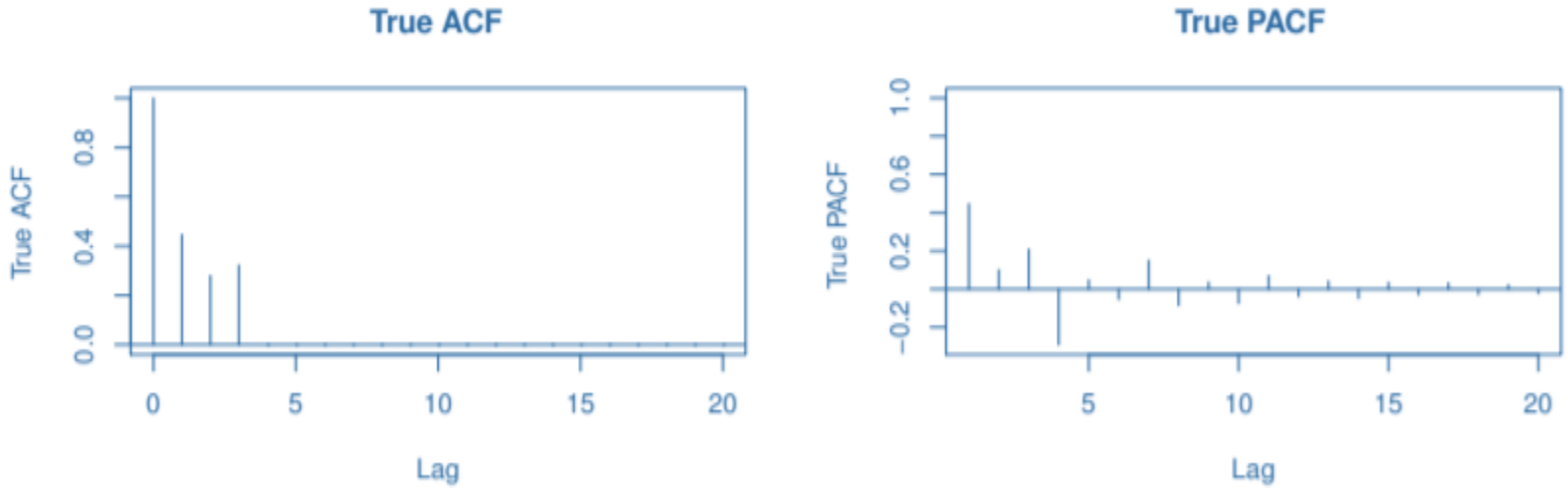
$$Y_t = \mu_t + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_q \varepsilon_{t-q}$$
$$\varepsilon_t \sim RB(0, \sigma^2)$$

Procesos de medias móviles de media q

Los procesos de medias móviles son la suma de procesos estacionarios por lo cual son siempre estacionarios.

Ajuste de los parámetros

[1] ¿Cómo establecer el valor de q?



Ajuste de los parámetros

[2] ¿Cómo establecer el valor de los parámetros θ_i?

Los parámetros están relacionados con las q raíces del polinomio y a través de otros polinomios y desarrollos:

$$\theta_q(z) = 1 + \theta_1 z + \theta_2 z^2 \ldots + \theta_q z^q$$

La función de autocorrelación parcial existe en un proceso MA(q) si las raíces se encuentran fuera del circulo unitario (condición de invertibilidad)

Ejercicio 1

Simular una serie de tiempo proveniente de un proceso MA(1)

```
n <-  300 # Número de periodos
theta  <-  c(1, 0.38, 1.55, 0.28, 0.75) # Parámetros
Mod(polyroot(theta)) # Raices del polinomio
y <- arima.sim(list(order=c(0,0,4), ma=theta[-1]), n=n, sd=sqrt(2.3))
ggtsdisplay(y)
```

1. ¿De cuántos rezagos depende la serie?
2. ¿Son las raíces todas con raíces mayores a 1?

Ejercicio 2

1. Generar una serie de un proceso MA(4)

2. Variar los valores de los parámetros a números positivos y negativos y verificar las raíces del polinomio

3. Variar los valores de los parámetros a números mayores que uno y menores que uno y verificar las raíces del polinomio

4. ¿Qué tipo de patrones del autocorrelograma y autocorrelograma parcial logra observar en estos procesos?

Procesos autorregresivos de orden p

El valor actual depende de forma lineal de los valores previos observados en la serie

Modelo medias móviles de orden 1

$$Y_t = \varphi_1 \mathrm{Y}_{\mathrm{t}-1} + \varepsilon_t$$
$$\varepsilon_t \sim RB(0, \sigma^2)$$

Modelo medias móviles de orden q

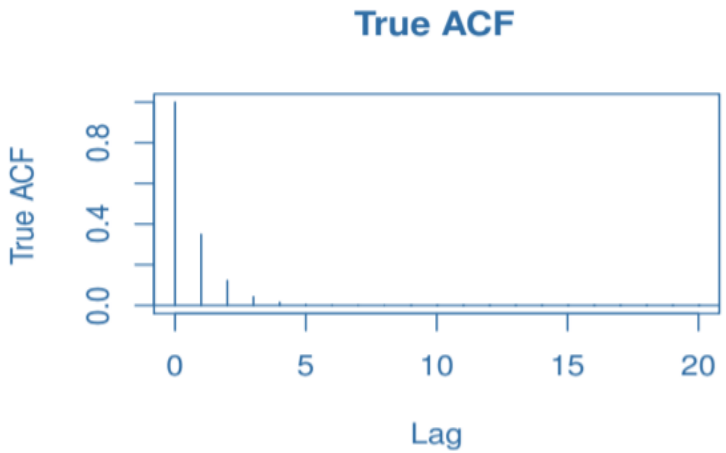
$$Y_t = \varphi_1 \mathrm{Y}_{\mathrm{t}-1} + \varphi_2 Y_{\mathrm{t}-2} + \cdots + \varphi_{t-p} Y_{t-p} + \varepsilon_t$$
$$\varepsilon_t \sim RB(0, \sigma^2)$$

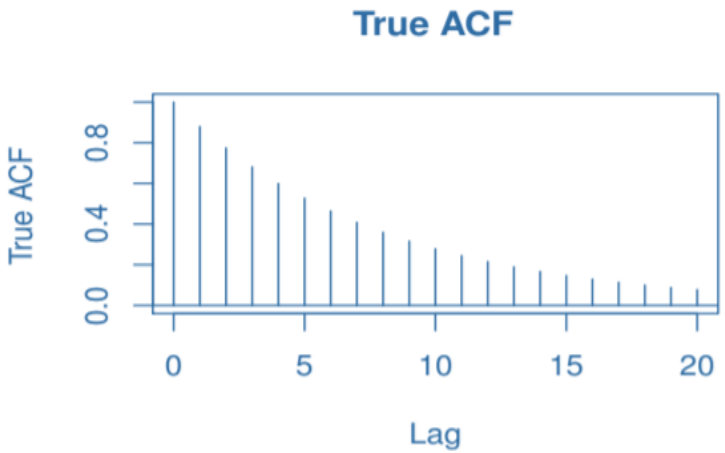
Procesos autorregresivos de orden p

A diferencia de los procesos de medias móviles, los procesos autoregresivos no siempre resultan siendo estacionarios, sin embargo el cumplimiento de esta hipótesis es importante en el proceso de predicción

Ajuste de los parámetros

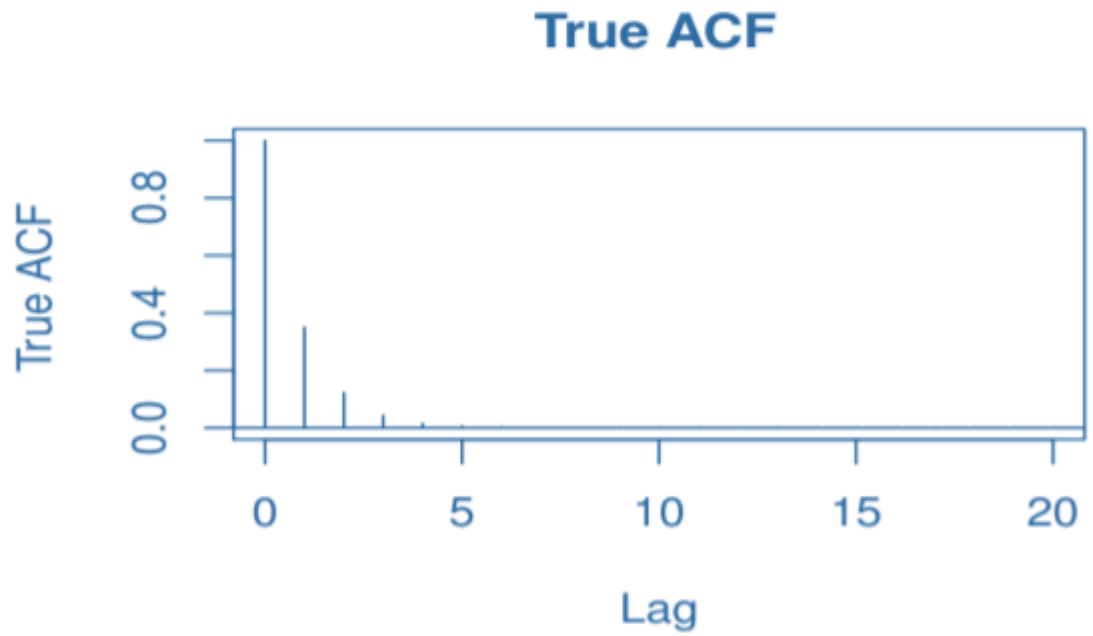
[1] ¿Cómo establecer el valor de q?





Ajuste de los parámetros

[1] ¿Cómo establecer el valor de q?



Garantizando la estacionaridad

Para que el proceso autoregresivo sea estacionario en covarianzas se debe asegurar que las p raíces del polinomio:

$$\theta_q(z) = 1 + \varphi_1 z + \varphi_2 z^2 \ldots + \varphi_p z^p$$

deben estar fuera el circulo unitario

Ejercicio 1

Simular una serie de tiempo proveniente de un proceso MA(1)

```
n <-  300 # Número de periodos
phi  <-  c(-1,-0.4,-0.4, -0.2, -0.7) # Parametros
Mod(polyroot(theta)) # Raices del polinomio
y <- arima.sim(list(order=c(4,0,0), ar=phi[-1]), n=n, sd=sqrt(2.3)
ggtsdisplay(y)
```

1. ¿De cuántos rezagos depende la serie?
2. ¿Son las raíces todas con raíces mayores a 1?

Ejercicio 2

1. Generar una serie de un proceso AR(4)

2. Variar los valores de los parámetros a números positivos y negativos y verificar las raíces del polinomio

3. Variar los valores de los parámetros a números mayores que uno y menores que uno y verificar las raíces del polinomio

4. ¿Qué tipo de patrones del autocorrelograma y autocorrelograma parcial logra observar en estos procesos?

Resumen

Proceso de medias móviles MA(q)

Los q primeros coeficientes del ACF son no nulos

Muchos de los coeficientes de la PACF no nulos que decrecen con el rezago como mezcla de sinusoidales o exponenciales

Proceso Autoregresivo AR(q)

Los q primeros coeficientes del FACF son no nulos

Muchos de los coeficientes de la ACF no nulos que decrecen con el rezago como mezcla de sinusoidales o exponenciales

Procesos Autorregresivos de promedios móviles ARMA

Procesos en que el comportamiento de la variable a través del tiempo depende de rezagos de sus propios valores y un término estocástico. Es el resultado de sumar varios procesos autorregresivos. El resultado es que el término de error es reemplazado por un proceso de medias móviles.

$$AR(p) + AR(q) = ARMA(p + q, \max(p, q))$$

Procesos Autorregresivos de promedios móviles ARMA

Modelo ARMA de orden 1,1

$$Y_t = c + \varphi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$
$$\varepsilon_t \sim RB(0, \sigma^2)$$

Modelo ARMA de orden p,q

$$Y_t - \varphi_1 Y_{t-1} - \cdots - \varphi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta \varepsilon_{t-1}$$
$$\varepsilon_t \sim RB(0, \sigma^2)$$

Procesos Autorregresivos de promedios móviles ARMA

En este modelo los errores no se tornan incorrelacionados, si no con una autocorrelación débil.

Estos modelos combinan la memoria a largo plazo de un proceso AR(p) con las propiedades de ruido débilmente correlacionado en los MA(q).

Procesos Autorregresivos de promedios móviles ARMA

Sus funciones de autocorrelación no tienen una forma especifica, es necesario ajustar varios modelos y usar criterios como AIC y BIC para estimar los valores de p y q

Ejercicio 1

Simular una serie de tiempo proveniente de un proceso ARMA(2,1)

```
n <-  300 # Número de periodos
theta <- c(1, 0.4)
phi  <-  c(-1,-0.4,-0.4)
y <- arima.sim(list(order=c(2,0,1), ma = theta[-1], ar=phi[-1]), n=n, sd=sqrt(2.3))
ggtsdisplay(y)
```

Modelos Autorregresivos de promedios móviles con tendencia estocástica ARIMA

Modelos que asumen que el comportamiento de la variable a través del tiempo depende de sus propios valores anteriores y un término estocástico y su media no es contante

Modelo ARIMA de orden p

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Error de una serie de tiempo

Medimos el error de una serie de tiempo como:

$$\varepsilon_t = Y_t - \widehat{Y_t}$$

Donde

$$\widehat{Y_t} = E(Y_t | Y_1, Y_2, \dots, Y_{t-1})$$

Error de una serie de tiempo

En un modelo para predecir los valores en una serie de tiempo se espera que los errores tengan un comportamiento de ruido blanco estacionario.

¿Cómo probar si una serie es de ruido blanco estacionario?

1. Comprobar que la media de los errores y hacer una gráfica para observar el comportamiento de los mismos
2. Hacer pruebas de no autocorrelación, una opción son las pruebas de Ljuing-Box y Durbin Watson

Prueba Ljung-Box (LB)

$$\begin{cases} H_o: \varepsilon_t \sim RB(0, \sigma^2) \\ H_1: \varepsilon_t \; no \; es \; ruido \; blanco \end{cases}$$

Box.test(r, lag = 25, type = "Ljung-Box")