Temperaturas medias mensuales de Barcelona (1995-2002)

Carla Martín Pérez

Series Temporales. Grado en Estadística. UGR

UNIVERSIDAD DE GRANADA

Fecha: 13 de enero de 2025

Índice

1. Introducción

Las series temporales se encuentran presentes en nuestra vida cotidiana en diversas formas y campos. Podemos ver ejemplos como en el ámbito de la salud, donde estas son fundamentales para el seguimiento de un paciente, o en meteorología, donde las temperaturas recogidas a lo largo del tiempo también forman parte de estas series temporales.

Por tanto, una definición de series temporales sería que son una sucesión de observaciones de una variables aleatoria en distintos instantes de tiempo.

En este trabajo se va a tratar de analizar detalladamente una serie temporal, con el objetivo de estudiar su comportamiento a lo largo del tiempo, identificando patrones, tendencias y posibles estacionalidades.

1.1. Información previa al estudio

Para hacer el estudio sobre una serie temporal se ha elegido una base de datos de datos.gob.es.

Esta base de datos, *datos_aire_Bc*, contiene información sobre las temperaturas medias mensuales del aire de la ciudad de Barcelona desde enero de 1780 hasta diciembre de 2023, proporcionada en grados centígrados (ºC). Esta base de datos tiene 4 variables:

- **Año**. Desde 1780 a 2023.
- **Mes**. Del 1 al 12, siendo el 1 enero y el 12 diciembre.
- **Nombre del mes**. Enero, febrero, ..., diciembre.
- **Temperatura**. Temperaturas en grados centrígrados (ºC)

Para realizar el estudio, más adelante, selecciono los datos a los años desde 1995 hasta 2002, para analizar la evolución de las temperaturas medias a lo largo de estos años.

2. Estudio de la serie temporal

Para empezar, voy a mostrar parte de los datos sobre los que voy a trabajar.

```
datos<-read.csv(file="datos_aire_Bc.csv", header=T, dec=".")

# Vemos las 6 primeras filas del archivo
head(datos)
```

Any	Mes	Desc_Mes	Temperatura
1780	1	Gener	6.7
1780	2	Febrer	7.3
1780	3	Març	11.6
1780	4	Abril	11.4
1780	5	Maig	16.3
1780	6	Juny	19.1

Al ser un número muy elevado de observaciones, voy a estudiar la serie temporal de la temperatura media de los meses desde el año 1995 hasta 2002.

Selecciono solo los datos de los meses pertenecientes a los años desde 1995 hasta 2002:

```
datos <- datos[datos$Any >= "1995" & datos$Any <= "2002", ]
head(datos)
```

	Any	Mes	Desc_Mes	Temperatura
2581	1995	1	Gener	8.6
2582	1995	2	Febrer	11.5
2583	1995	3	Març	10.9
2584	1995	4	Abril	14.0
2585	1995	5	Maig	17.2
2586	1995	6	Juny	19.6

Se puede ver la estructura de las datos:

```
# Ver estructura de los datos
str(datos)
```

```
'data.frame':   96 obs. of  4 variables:
 $ Any        : int  1995 1995 1995 1995 1995 1995 1995 1995 1995 1995 ...
 $ Mes        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Desc_Mes   : chr  "Gener" "Febrer" "Març" "Abril" ...
 $ Temperatura: num  8.6 11.5 10.9 14 17.2 19.6 25 23.3 18.7 18.3 ...
```

Como ya se mencionó anteriormente, existen 4 columnas que se tratan de las variables, *Any*, *Mes*, *Desc_Mes*, *Temperatura*, y 96 filas, que se refiere al número de temperaturas medias de los meses de cada uno de los años (1995-2002) recogidas en el archivo.

También podemos ver un resumen de estos datos:

```
summary(datos)
```

```
      Any            Mes          Desc_Mes          Temperatura   
 Min.   :1995   Min.   : 1.00   Length:96          Min.   : 6.80  
 1st Qu.:1997   1st Qu.: 3.75   Class :character   1st Qu.:11.00  
 Median :1998   Median : 6.50   Mode  :character   Median :15.15  
 Mean   :1998   Mean   : 6.50                      Mean   :15.77  
 3rd Qu.:2000   3rd Qu.: 9.25                      3rd Qu.:20.75  
 Max.   :2002   Max.   :12.00                      Max.   :25.00  
```

Con este resumen de los datos se puede ver la temperatura mensual mínima alcanzada, siendo esta de 6.8ºC, y la máxima, 25ºC, desde 1995 hasta 2002. El resto de variables no las observaremos ya que trabajaremos sobre las temperaturas.

Y para empezar a trabajar sobre ellas convertiremos la variable *Temperatura* de nuestros datos en una serie temporal.

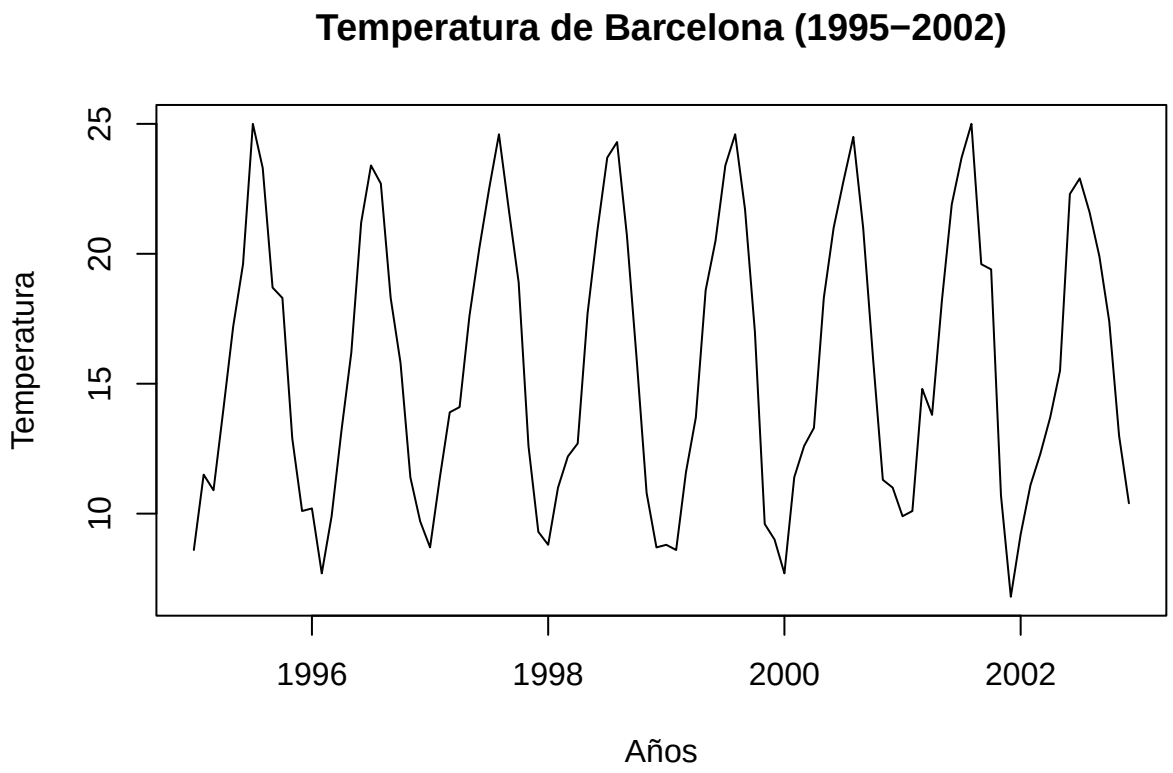
```
temp<-ts(datos$Temperatura, start=1995, frequency = 12)

# Vemos que clase es la variable que hemos creado
class(temp)
```

```
[1] "ts"
```

Al devolvernos **"ts"** con la función *class*, comprobamos que se ha creado un objeto (variable) de clase "ts", **Time-Series**, correctamente.

Podemos ver esta serie temporal gráficamente:

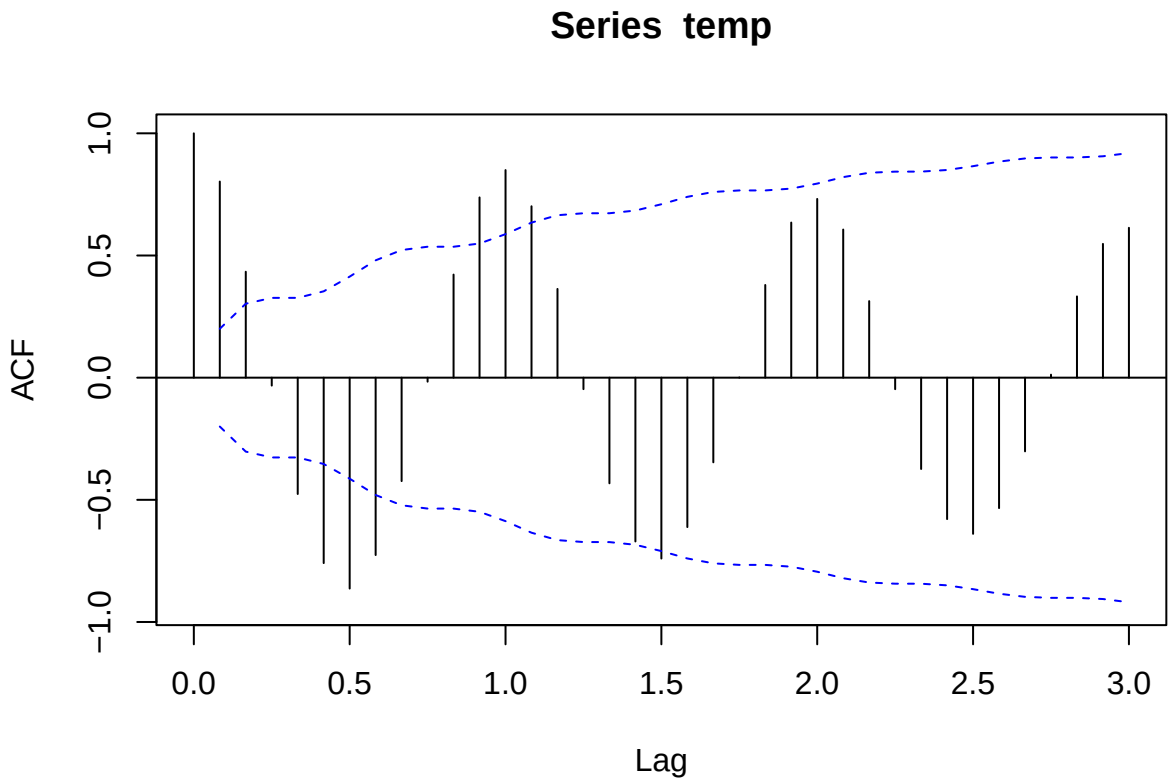
```
plot(temp, ylab = "Temperatura", xlab = "Años", main = "Temperatura de Barcelona (1995-2002)")
```



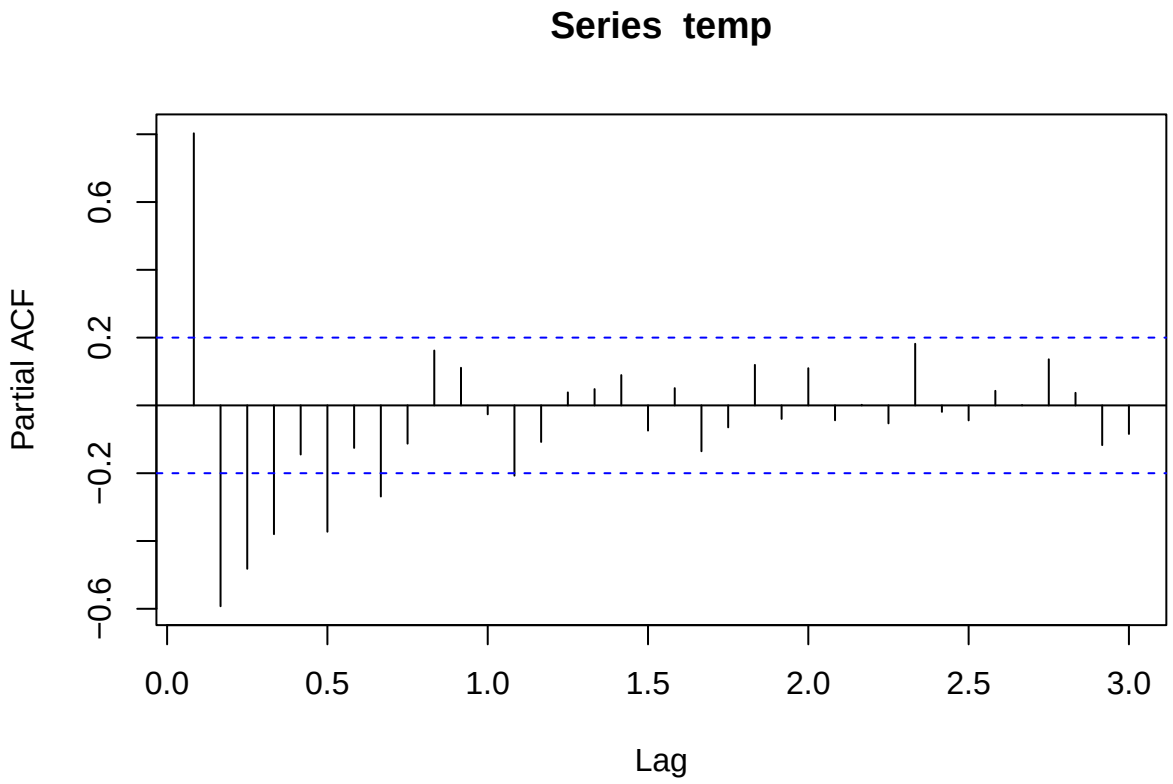
A simple vista, podemos ver diferentes temperaturas a lo largo de las años, unas más alta que otras, viendo que en los meses de verano las temperaturas crecen, y en los meses de invierno las temperaturas disminuyen, pero no parece haber una tendencia lineal clara hacia un aumento o descenso de las temperaturas. Tampoco se podría asegurar la existencia de un patrón estacional a simple vista.

Para poder analizar mejor la serie mostraré las funciones de autocorrelación, FAC, y autocorrelación parcial, FACP, y para dibujarlas elegimos hacerlo para un retardo máximo de 36, ya que se sugiere tomar un valor para un retardo máximo múltiplo de del periodo estacional, por lo que, en este caso, $\frac{T}{3} = 32$, el retardo máximo estará entre 36 y 24 que son los más cercano a este valor, y elegiremos 36.

```
acf(temp, lag.max = 36, ci.type = "ma")
```




```
pacf(temp, lag.max= 36)
```



En la función de autocorrelación de la serie, el primer valor en el lag 0 es siempre 1, porque está correlacionado consigo mismo, pero a partir del lag 1, aunque algunas sobre pasan las bandas del intervalo de confianza, se puede observar un leve decrecimiento. Y en la función de autocorrelación parcial se puede ver que los lags significativos son los encontrados al principio y alguno aislado más adelante.

Lo siguiente es el test de Dickey-Fuller, este test es imortante para comprobar la presencia de raíces unitarias, es decir, si su comportamiento está dominado por una tendencia no estacionaria, por lo que la hipótesis nula, H_0, consiste en que hay una raíz unitaria.

```
library(fUnitRoots)
adfTest(temp, lags = 1)
```

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
```

```
  STATISTIC:
    Dickey-Fuller: -1.7423
  P VALUE:
    0.08055

Description:
 Mon Mar  3 18:20:00 2025 by user: carla
```

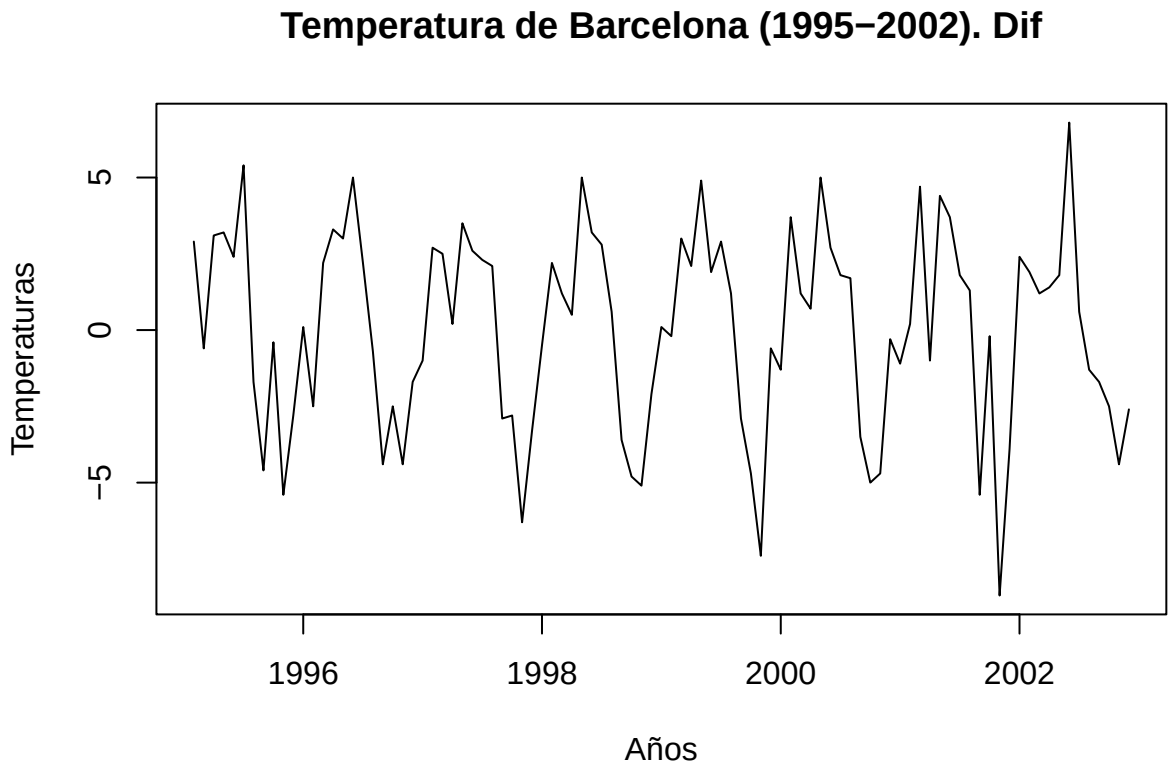
El p-valor obtenido es 0.08055 con lo que para $\alpha = 0{,}05$, no podemos rechazar la hipótesis nula consistente en que hay una raíz unitaria, es decir, no se puede rechazar que su comportamiento esté dominada por una tendencia no estacionaria.

Tras este resultado, lo siguiente que se debe hacer es diferenciar la serie.

```
dtemp<-diff(temp)
```

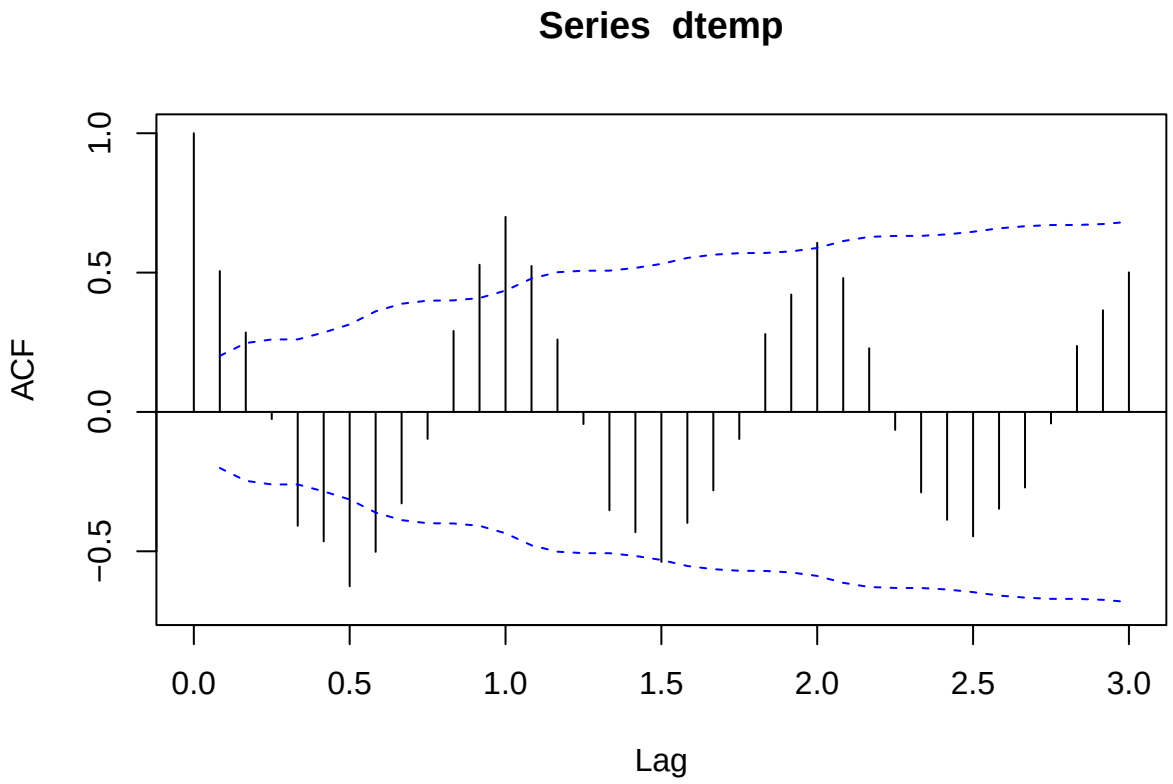
Si dibujamos la serie diferenciada, tenemos:

```
plot(dtemp, ylab = "Temperaturas", xlab = "Años", main = "Temperatura de Barcelona (1995-2002).
```

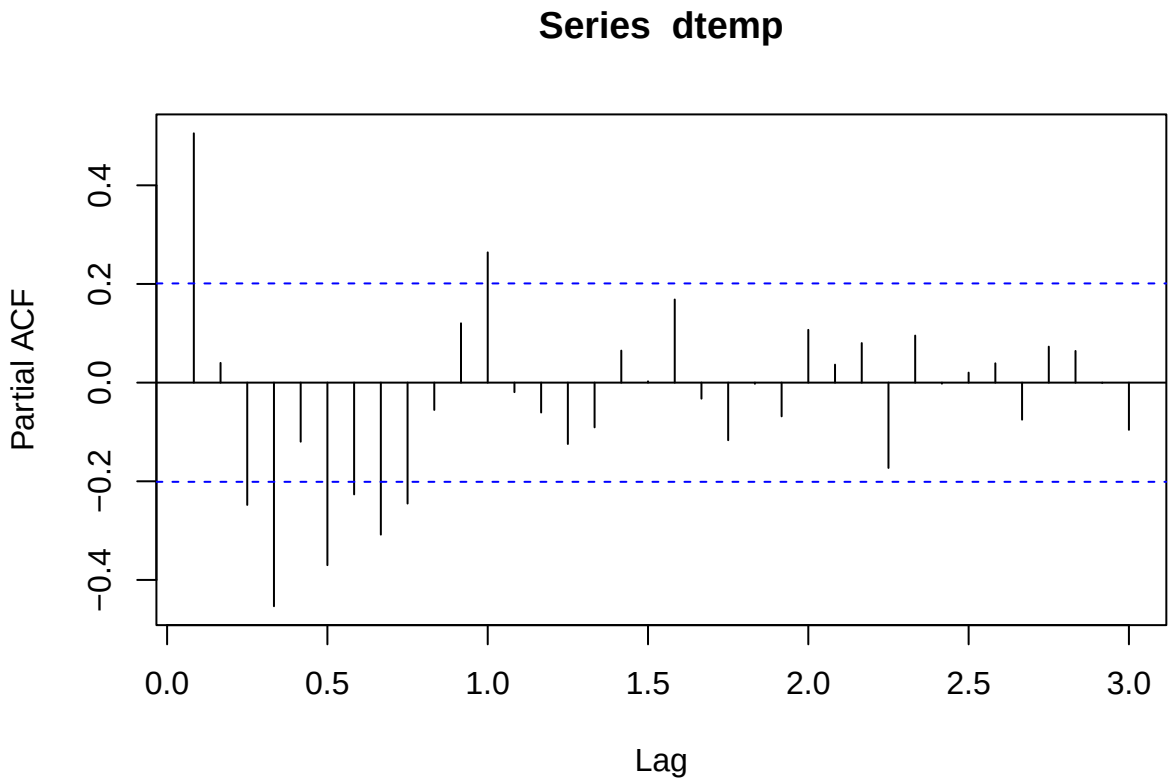


La gráfica de la serie diferenciada nos permite decir que la serie podría ser estacionaria, ya que mantiene una variabilidad constante y se va moviendo alrededor de un nivel medio, 0. Por lo que después de estas conclusiones, para poder confirmarlo, vuelvo a dibujar las funciones de autocorrelación y autocorrelación parcial.

```
acf(dtemp, lag.max = 36, ci.type = "ma")
```




```
pacf(dtemp, lag.max = 36)
```

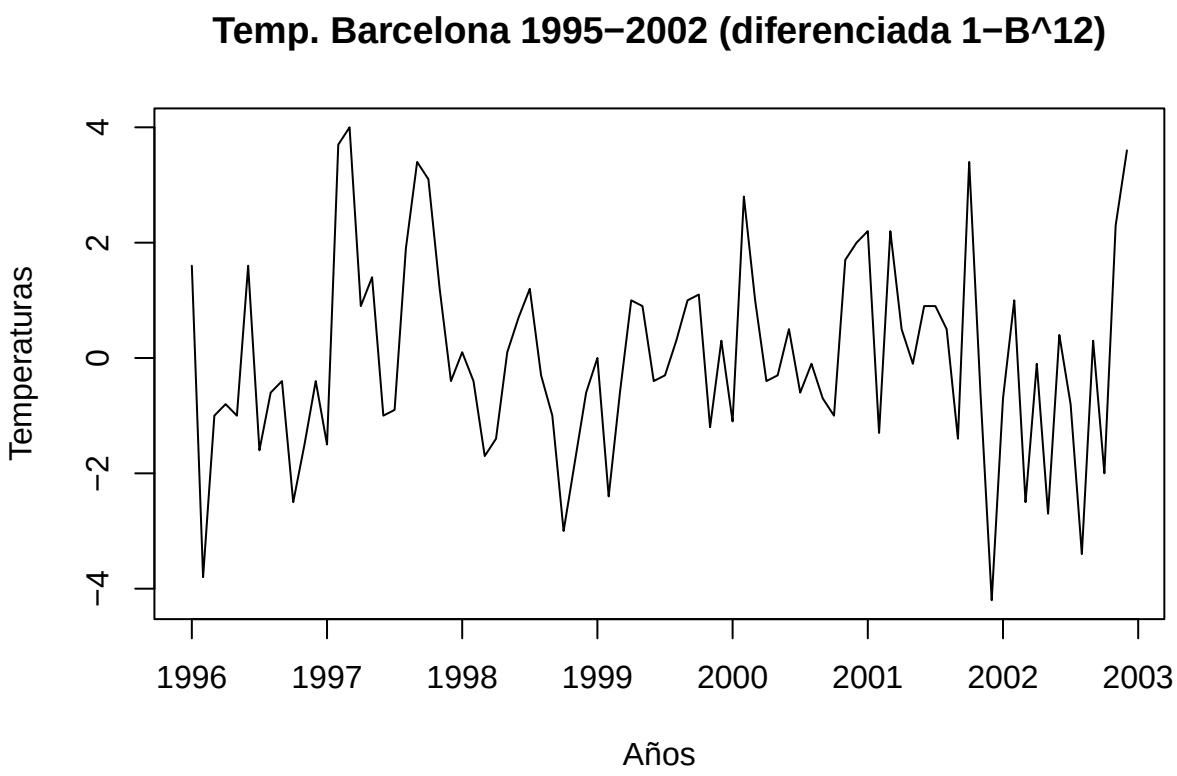


En la función de autocorrelación podemos ver un decrecimiento, aunque con algunos retardos significativos, entre ellos, el primer y segundo retardo estacional (correspondientes al duodécimo y vigésimo cuarto retardo, por ser datos mensuales). En la función de autocorrelación parcial también notamos un decrecimiento con retardos significativos, en este caso, también es significativo el primer retardo estacional, por lo que en ambas funciones nos encontramos esta característica. Este hecho, nos hace pensar que deberiamos realizar una diferenciación estacional, es decir, en lugar de diferenciar la serie mediante $(1-B)$, se diferenciaría primero usando una diferenciación estacional, en este caso $(1-B^{12})$.

```
dstemp<-diff(temp, lag=12, differences = 1)
```

Podemos ver la gráfica de esta serie diferenciada estacionalmente:

```
plot(dstemp, ylab = "Temperaturas", main="Temp. Barcelona 1995-2002 (diferenciada 1-B^12)",xlab
```



Esta serie parece ser estacionaria en cuanto varianza y media, pero para poder confirmarlo realizo el test de Dickey-Fuller.

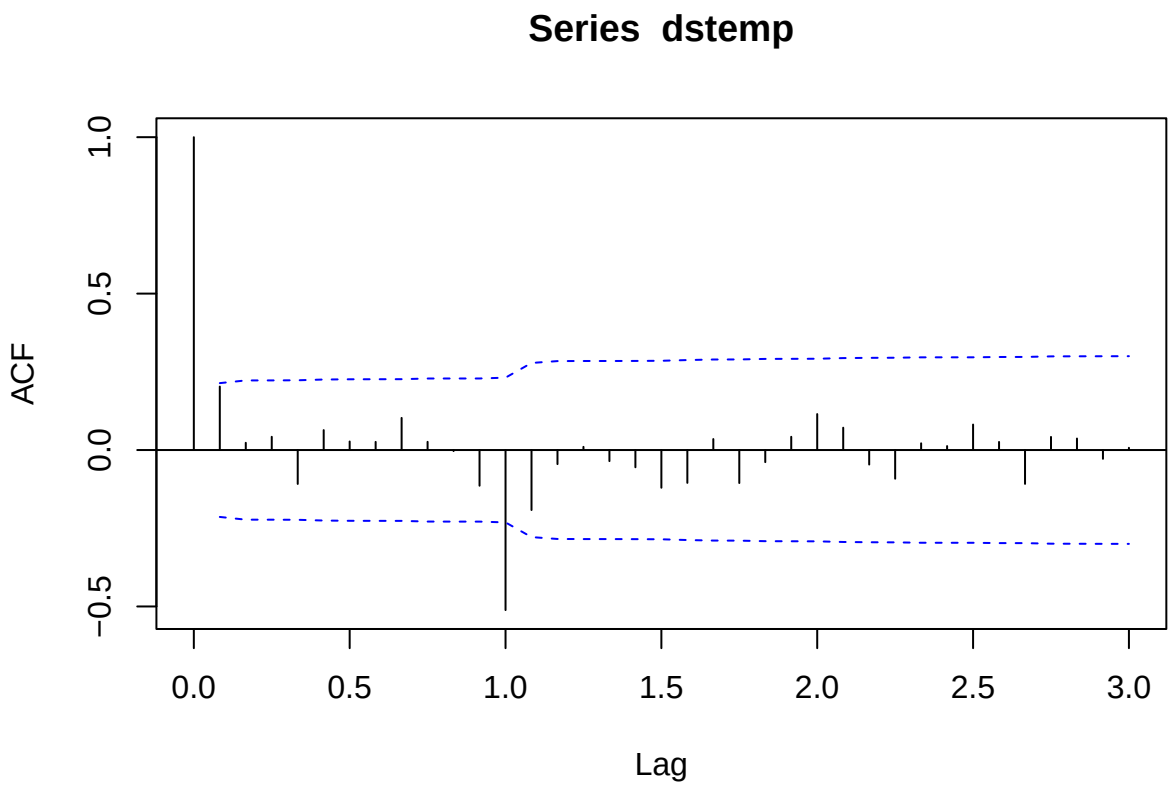
```
adfTest(dstemp, lags = 1)
```

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 1
  STATISTIC:
    Dickey-Fuller: -5.4809
  P VALUE:
    0.01

Description:
 Mon Mar  3 18:20:09 2025 by user: carla
```

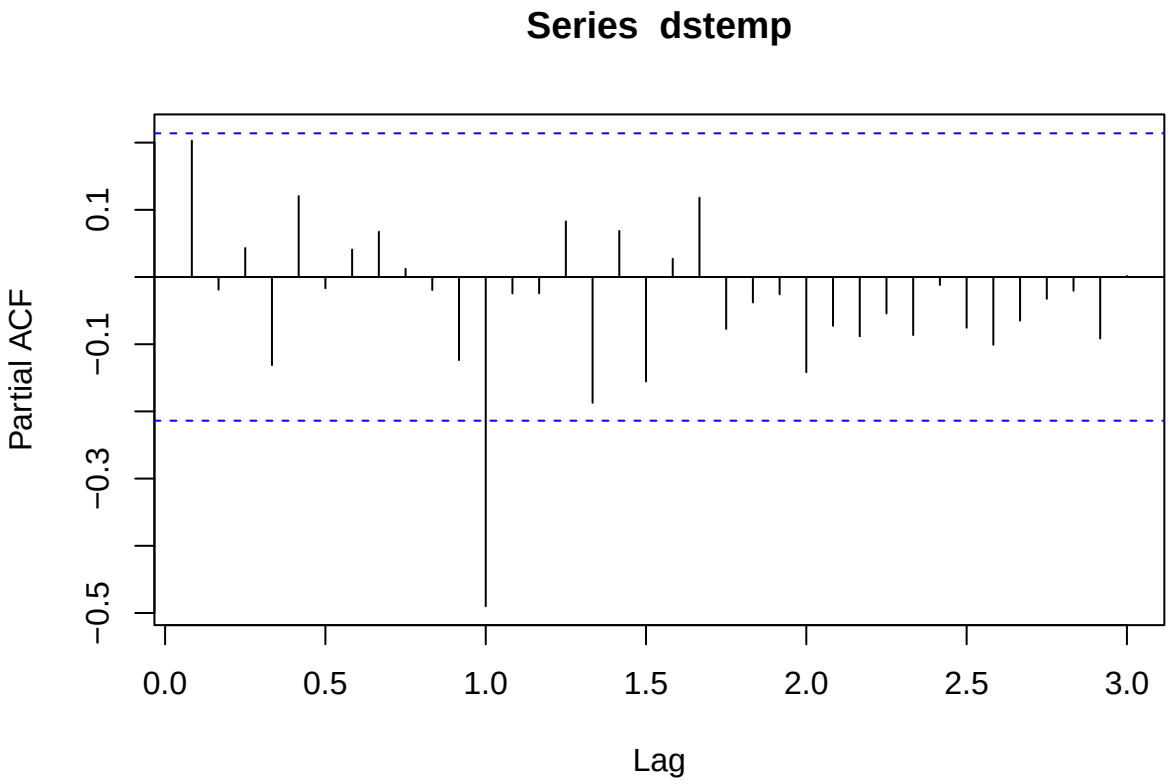
Como el p-valor obtenido es 0.01 con lo que para $\alpha = 0{,}05$, podemos rechazar la hipótesis nula consistente en que hay una raíz unitaria, es decir, se puede rechazar que su comportamiento esté dominada por una tendencia no estacionaria.

Las gráficas de FAC y FACP de esta serie son las siguientes:

```
acf(dstemp,lag.max=36,ci.type="ma")
```




```
pacf(dstemp,lag.max=36)
```



Fijándonos en los gráficos anteriores podemos obtener que modelo se debe ajustar. Voy a ajustar un modelo $SARIMA(0,0,0)(0,1,1)_{12}$, y otro $SARIMA(0,0,0)(1,1,0)_{12}$, pero antes de empezar a ajustar estos modelos mencionaré de que se compone un modelo **SARIMA**.

Un modelo $SARIMA(p,d,q)(P,D,Q)_s$, *Seasonal AutoRegressive Integrated Moving Average*, es una extensión del modelo ARIMA que incorpora componentes estacionales. Este modelo está compuesto por dos conjuntos de parámetros, los no estacionales, (p, d, q), donde:

- p: orden de la parte autorregresiva (AR). En nuestro caso, p = 0, ya que no hay evidencia de un término autorregresivo no estacional, es decir, el primer retardo en la FACP no es significativo, siendo los retardos que le continúan no significativos.
- d: grado de diferenciación. Será d = 0, porque no se realiza una diferenciación (1 - B) en nuestros datos.
- q: orden de la parte de medias móviles (MA). q = 0, el modelo no utiliza la relación entre los errores pasados para predecir los valores futuros.

Y los estacionales , $(P,D,Q)_s$, donde:

- P: orden autoregresivo estacional. Es P = 1, el modelo incluye una dependencia de los valores de la serie temporal registrados en el mismo punto de tiempo en ciclos estacionales anteriores. En la PACF se encuentra que el primer retardo estacional es significativo.

- D: grado de diferenciación estacional. D = 1 porque se ha diferenciado estacionalmente la serie.
- Q: orden de las medias móviles estacional. Q = 1, el modelo considera la influencia de los errores de predicción en ciclos estacionales anteriores. En la ACF se muestra un pico significativo en el primer retardo estacional.
- s: periodo de estacionalidad (número de observaciones por ciclo estacional). En mi caso, s = 12, ya que los datos están recogidos mensualmente.

Se ajustarán los modelos mencionados:

SARIMA(0, 0, 0)(0, 1, 1)_12

```
ms1 <- arima(temp, order = c(0, 0, 0), seasonal = list(order = c(0, 1, 1), period = 12), metho
ms1
```

```
Call:
arima(x = temp, order = c(0, 0, 0), seasonal = list(order = c(0, 1, 1), period = 12),
    method = "ML")

Coefficients:
         sma1
      -0.9998
s.e.   0.2526

sigma^2 estimated as 1.339:  log likelihood = -143.9,  aic = 291.81
```

Para saber si los coeficientes son significativos en este primer modelo, vemos si en el IC al 95 %, $Coeficiente \pm 1{,}96 \times s.e.$, no se encuentra el 0:

- Para **sma1**:

IC al 95 %: $-0{,}9998 \pm 1{,}96 \times 0{,}2526 = [-1{,}4949, -0{,}5047]$

Por tanto, como $0 \notin [-1{,}4949, -0{,}5047]$, el coeficiente *sma1* es significativamente distinto de 0.

SARIMA(0, 0, 0)(1, 1, 0)_12

```
ms2 <- arima(temp, order = c(0, 0, 0), seasonal = list(order = c(1, 1, 0), period = 12), metho
ms2
```

```
Call:
arima(x = temp, order = c(0, 0, 0), seasonal = list(order = c(1, 1, 0), period = 12),
    method = "ML")

Coefficients:
         sar1
      -0.6369
s.e.   0.0866

sigma^2 estimated as 1.78:  log likelihood = -146.54,  aic = 297.07
```

Para saber si los coeficientes son significativos en este segundo modelo:

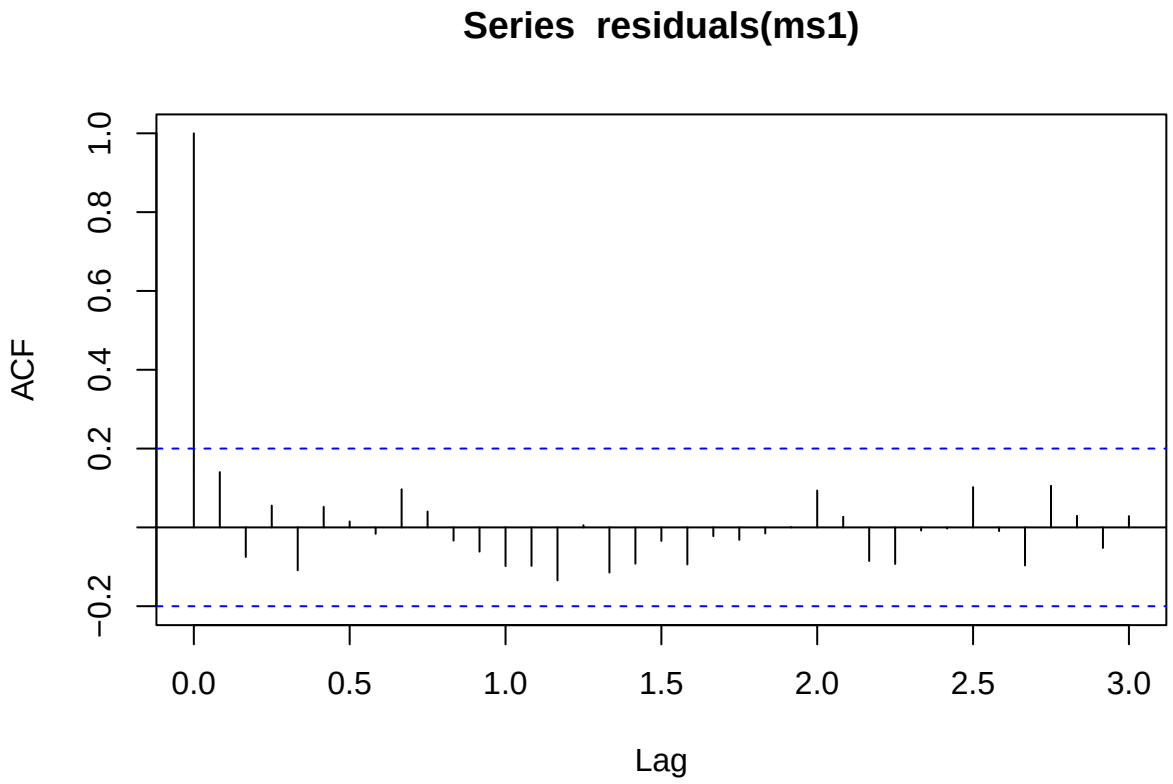
- Para **sar1**:

IC al 95 %: $-0{,}6369 \pm 1{,}96 \times 0{,}0866 = [-0{,}8066, -0{,}4672]$

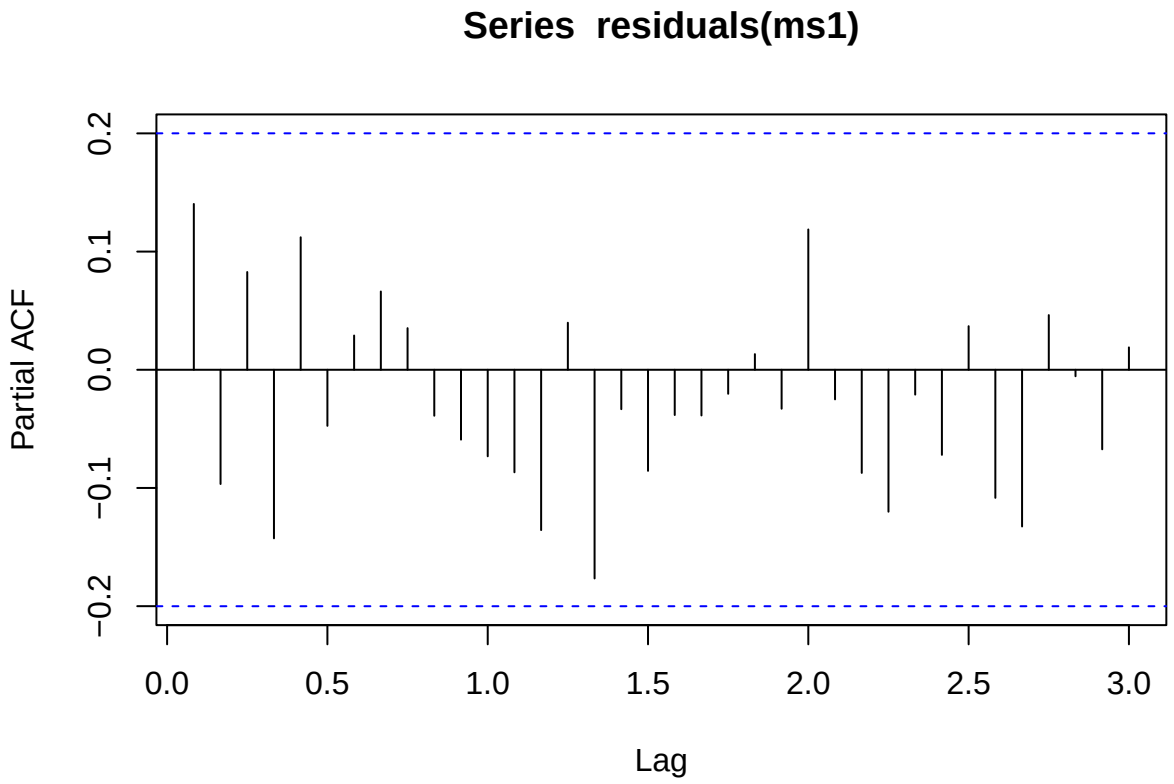
Por tanto, como $0 \notin [-0{,}8066, -0{,}4672]$, el coeficiente *sar1* es significativamente distinto de 0.

En ambas salidas hemos obtenido la métrica **AIC**, Criterio de Información de Akaike, por su siglas en inglés, de cada modelo. Esta métrica nos devuelve qué tan bien se ajusta un modelo a los datos. Un menor AIC indica un mejor modelo, ya que se ofrece un mejor balance entre ajuste y simplicidad. En mi caso, en el primer modelo (**ms1**) tenemos, **AIC = 291.81**, frente al valor de este en el segundo modelo (**ms2**), **AIC = 297.07**. Por lo que el primer modelo será mejor que el segundo.

```r
acf(residuals(ms1), lag.max = 36)
```




```
pacf(residuals(ms1), lag.max = 36)
```



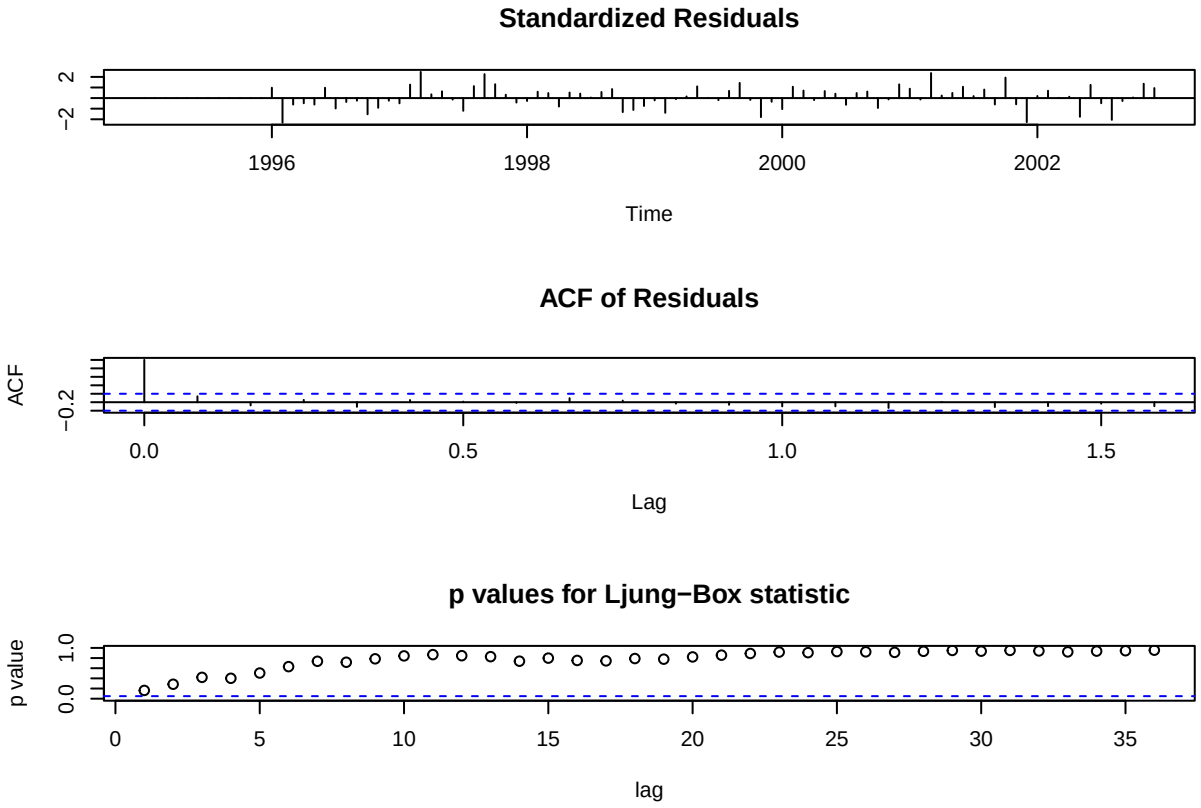
Los valores de la función de autocorrelación y autocorrelación parcial no son singnificativamente distintos de cero, como concluimos anteriormente cuando mencionamos que $0 \notin [-1{,}4949, -0{,}5047]$, por lo que el parámetro **sma1** es significativamente distinto de cero.

Ahora realizaré el **test de Ljun-Box**. Este test es un herramiento que nos ayuda a evaluar si los residuos de un modelo temporal son independientes, es decir, si son ruido blanco. Por lo que las hipotesis de este test son:

H_0 : Los residuos son independiente (ruido blanco),

H_1 : Los residuos tienen autocorrelación significativa (no son ruido blanco)

```
tsdiag(ms1, gof=36)
```



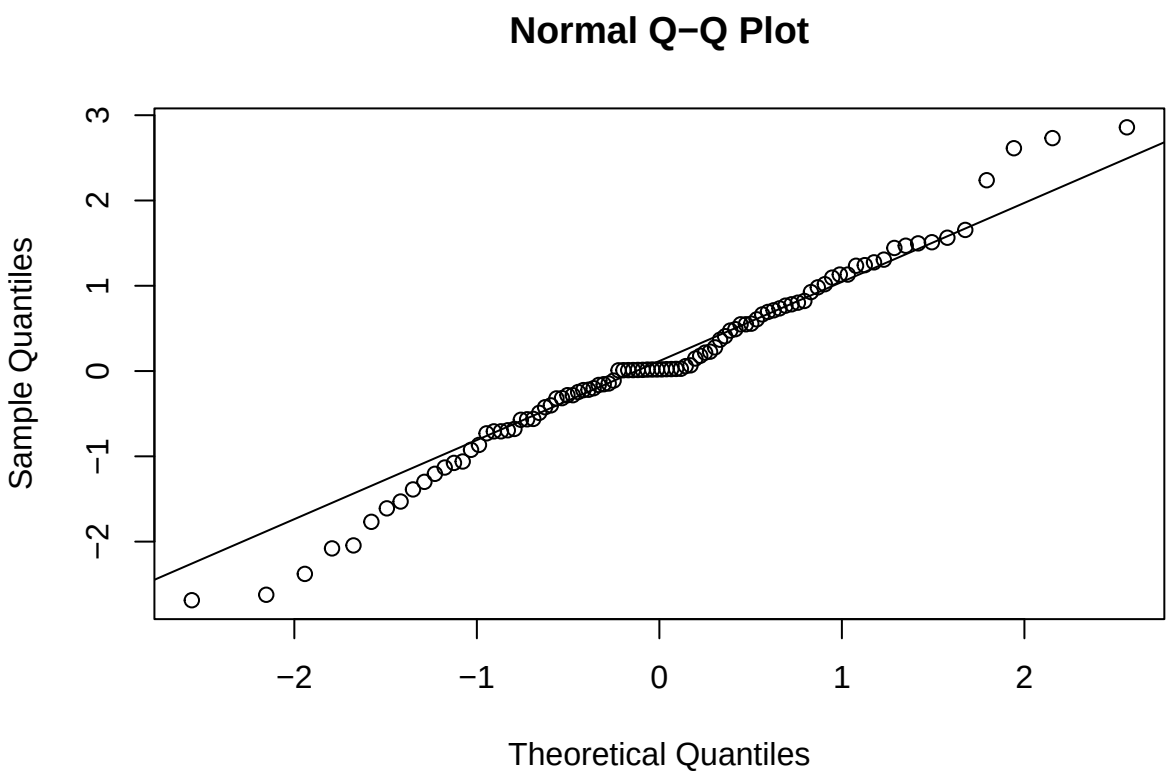

```
Box.test(residuals(ms1),lag = 36, type = "Ljung-Box", fitdf = 2)
```

```
    Box-Ljung test

data:  residuals(ms1)
X-squared = 22.924, df = 34, p-value = 0.9253
```

En este test hemos obtenido un p-valor $= 0.9253 > 0.05$, con lo que no podemos rechazar la hipótesis nula, H_0 : Los residuos son independiente (ruido blanco),, es decir, el modelo $SARIMA(0,0,0)(0,1,1)_{12}$ es podría ser adecuado.

Estudiamos también la normalidad de los residuos de este modelo a través de la siguiente gráfica:

```
qqnorm(residuals(ms1))
qqline(residuals(ms1))
```



Los residuos del modelo parecen seguir una distribución aproximadamente normal, salvo por algunos puntos que se alejan de la diagonal, especialmente al principio y al final de ella, pero aún así no se puede rechazar la **normalidad**.

Para confirmar la normalidad de los residuos del modelo, voy a realizar el **test de Shapiro.Wilk**:

```
shapiro.test(residuals(ms1))
```

```
    Shapiro-Wilk normality test

data:  residuals(ms1)
W = 0.98233, p-value = 0.2235
```

Al obtener un p-valor $= 0.2235 > 0.05$, no podemos rechazar la hipótesis nula, es decir, no hay evidencias estadísticas para rechazar que los residuos siguen una distribución normal.

También calcularé las correlaciones de los estimadores:

```
AA = ms1$var.coef
dd = dim(AA)
B = diag(AA)
CC = diag(dd[1])
for(i in 1:dd[1]){
  CC[i,i] = 1/sqrt(B[i])
}

CC%*%AA%*%CC
```

```
     [,1]
[1,]    1
```

Al tener un único parámetro en el modelo, **sma1**, nos devuelve 1, ya que no puede estar correlacionado con otro.

El modelo ajustado a los datos es:

$$(1)(1 - B^{12})X_t = (1)(1 - 0{,}9998B^{12})\epsilon_t$$

Para concluir el estudio, realizo la predicción de la serie en el año siguiente:

```
library(forecast)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

```
datos_predicción <- forecast(ms1, h=12)
datos_predicción
```

```
         Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
Jan 2003       8.987499  7.415029 10.55997  6.582613 11.39238
Feb 2003      10.349999  8.777529 11.92247  7.945113 12.75488
Mar 2003      12.274999 10.702529 13.84747  9.870113 14.67988
Apr 2003      13.562498 11.990028 15.13497 11.157613 15.96738
May 2003      17.412498 15.840028 18.98497 15.007612 19.81738
Jun 2003      20.949998 19.377527 22.52247 18.545112 23.35488
Jul 2003      23.424997 21.852527 24.99747 21.020111 25.82988
Aug 2003      23.824997 22.252527 25.39747 21.420111 26.22988
```

```
Sep 2003      20.199998 18.627527 21.77247 17.795112 22.60488
Oct 2003      17.337498 15.765028 18.90997 14.932612 19.74238
Nov 2003      11.537499  9.965028 13.10997  9.132613 13.94238
Dec 2003       9.374999  7.802529 10.94747  6.970113 11.77988
```

Como curiosidad, ya que la base de datos utilizada llega hasta 2023, podemos comparar estas predicciones con los datos reales:

```
datos_iniciales <-read.csv(file="datos_aire_Bc.csv", header=T, dec=".")

datos_reales <- datos_iniciales[datos_iniciales$Any == "2003", ]
datos_reales
```

	Any	Mes	Desc_Mes	Temperatura
2677	2003	1	Gener	7.9
2678	2003	2	Febrer	7.1
2679	2003	3	Març	12.1
2680	2003	4	Abril	14.0
2681	2003	5	Maig	18.0
2682	2003	6	Juny	25.6
2683	2003	7	Juliol	26.1
2684	2003	8	Agost	28.5
2685	2003	9	Setembre	20.8
2686	2003	10	Octubre	15.2
2687	2003	11	Novembre	12.7
2688	2003	12	Desembre	8.8