Análisis de Series de Tiempo

Carrera de Especialización en Inteligencia Artificial

Clase 4

Ing. Magdalena Bouza, Ing. Carlos German Carreño Romano

Temas

- Repaso ARIMA - operador backshift
- objeto de resultados ARIMA
- Power transform, log
- Diagnósticos
- ejemplos de función de autocorrelación y autocorrelación parcial
- (S)ARIMA

ARIMA

Modelo ARIMA

Diremos que *{Yt}* sigue un modelo ARIMA si $W_t = \nabla^d Y_t$ es un proceso ARMA(p,q) estacionario. Diremos en este caso que *Yt* sigue un proceso ARIMA(p,d,q).

Operador de *backshift*: el operador de backshift (B) opera sobre el índice temporal de la serie de tiempo y la desplaza en una unidad de tiempo:

$$BY_t = Y_{t-1}$$

El operador B es lineal: $B(aY_t + bX_t + c) = aY_{t-1} + bX_{t-1} + c$

En general escribimos que $B^d Y_t = Y_{t-d}$

ARIMA + backshift

MA(q):

$$\begin{aligned} Y_t &= e_t - b_1 e_{t-1} - \ldots - b_q e_{t-q} \\ &= e_t - b_1 B e_t - \ldots b_q B^q e_t \quad \Rightarrow Y_t = b(B)e_t \\ &= (1 - b_1 B - \ldots - b_q B^q) e_t \end{aligned}$$

ARMA(p,q)

$$a(B)Y_t = b(B)e_t$$

AR(p)

$$\begin{aligned} Y_t &= e_t + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} \\ Y_t - a_1 Y_{t-1} - \ldots - a_p Y_{t-p} &= e_t \\ Y_t - a_1 B Y_t - \ldots - a_p B^p Y_t &= e_t \quad \Rightarrow a(B)Y_t = e_t \\ (1 - a_1 B - \ldots - a_p B^p) Y_t &= e_t \end{aligned}$$

ARIMA(p,d,q)

$$a(B)(1-B)^d Y_t = b(B)e_t$$

Sobre la estimación del orden de integración

Al estimar el orden *d* en el modelo ARIMA(p,d,q) se debe tener cuidado de no diferenciar de más.

Si bien la diferencia de cualquier serie estacionaria sigue siendo estacionaria, si diferenciamos más de lo necesario estamos introduciendo correlaciones innecesarias en los datos, complicando el modelado.

Además sobrediferenciar lleva a modelos no invertibles, los cuales traen problemas a la hora de estimar los parámetros del modelo

Otras transformaciones

Transformación logarítmica

¿Cuándo usar la transformación logarítmica?

- Cuando observamos que la varianza del proceso parece aumentar con el tiempo.
  En particular si $\mathbb{E}[Y_t] = \mu_t$ y $\sqrt{var(Y_t)} = \mu_t \sigma$, luego $\mathbb{E}[\log(Y_t)] \approx \log(\mu_t)$ y $var(\log(Y_t)) \approx \sigma^2$
- Si *Yt* tiene cambios porcentuales relativamente estables entre un instante de tiempo y otro, y supongamos que $Y_t = (1 + X_t)Y_{t-1}$. Luego tomando el log

  $$\log(Y_t) = \log((1 + X_t)Y_{t-1}) = log(1 + X_t) + log(Y_{t-1}) \rightarrow \log(Y_t) - \log(Y_{t-1}) = \log\left(\frac{Y_t}{Y_{t-1}}\right) = \log(1 + X_t)$$

  Si además suponemos que *Xt* está acotado, |*Xt*|<0.2, sucede que $log(1 + X_t) \approx X_t$ y $\nabla(\log Y_t) \approx X_t$ va a ser relativamente estable y posiblemente se encuentre bien modelada por un poc. estacionario.

Diagnóstico de modelos

Diagnóstico de los modelos

Se basa en determinar la bondad de la estimación del modelo, y en caso de tener un mal ajuste proponer modificaciones apropiadas.

Vamos a ver dos enfoques complementarios:

- Análisis residual
- Análisis de modelos sobreparametrizados

1. Análisis de residuos

Al igual que en los problemas de regresión, vamos a llamar residuo a la diferencia entre el valor verdadero y el estimado por el modelo:

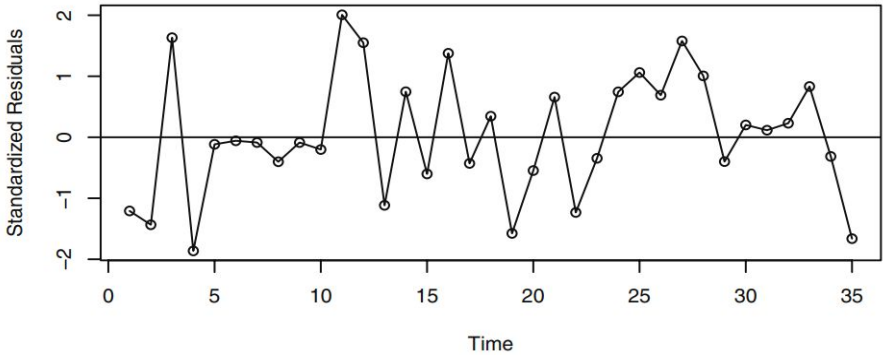
$$\hat{e}_t = Y_t - \hat{Y}_t$$

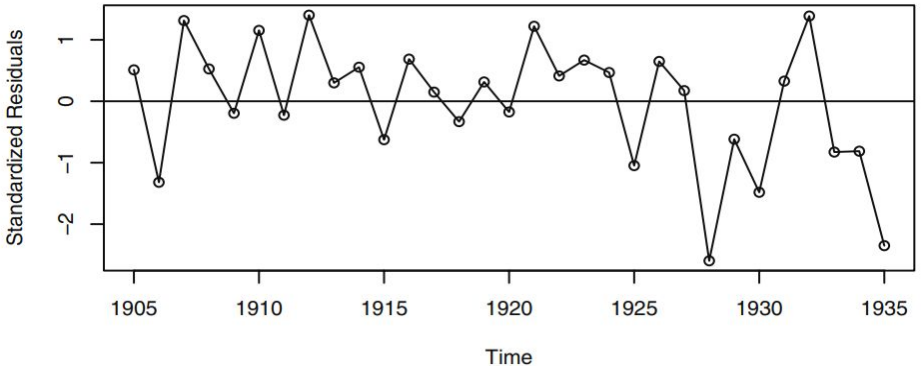
Si el modelo se encuentra bien estimado, los residuos deberían tener aproximadamente las propiedades del ruido blanco.

Las desviaciones de esta propiedad pueden ayudarnos a corregir el modelo.

Gráfico de residuos

El primer testeo de diagnóstico consiste en graficar los residuos obtenidos a lo largo del tiempo.

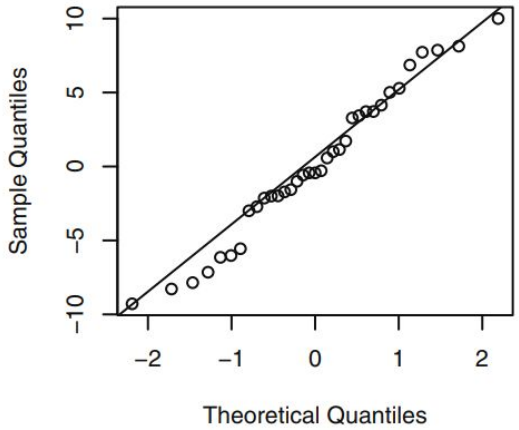


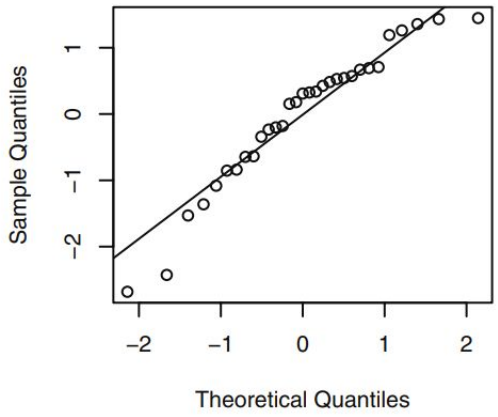


Normalidad de los residuos

Una buena forma de analizar la normalidad de los residuos es mediante un QQplot.

También pueden analizarse tests estadísticos como el Shapiro-Wilk (recordar que si la cantidad de muestras es muy grande este camino puede no ser recomendable).





Autocorrelación de los residuos

Para analizar la independencia de los residuos podemos estimar su función de autocorrelación.

Idealmente, para tamaños grandes de muestra, las estimaciones de las correlaciones seguir una distribución normal de media 0 y varianza 1/n. Sin embargo esto no resulta del todo cierto para lags pequeños j, k, donde la varianza puede resultar significativamente menor y estar altamente correlacionados

Test de Ljung-Box

Además de analizar las autocorrelaciones de los residuos para lags individuales, es bueno tener una métrica que contemple las magnitudes de estas autocorrelaciones en conjunto.

Box y Pierce propusieron el estadístico $Q = n(\hat{r}_1^2 + \hat{r}_2^2 + \ldots + \hat{r}_k^2)$. Mostraron que si los órdenes *p* y *q* del ARMA están bien estimados, y *n* es grande, $Q \approx \chi^2_{k-p-q}$.
El problema que es la dist. asintótica está basada en un teorema límite.

Ljung y Box demostraron que esta dist. no se cumple para para tamaños comunes de muestras. Propusieron $Q_* = n(n+2)(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-1} + \ldots + \frac{\hat{r}_k^2}{n-1})$ que se asemeja mucho más a la distribución chi2.

En ambos casos el test asociado es H0: “Los residuos están descorrelacionados”.
Si el p-valor asociado al test resulta superior a un ublar (ej 0.05) no tengo evidencia suficiente para rechazar la hip. que los residuos están descorrelacionados.

2. Overfitting y redundancia de parámetros

Luego de ajustar un modelo que nos parece razonable, ajustamos un algo más general (de más parámetros), que contenga al modelo propuesto como un caso particular. Por ejemplo, si queremos valildar un AR(2), podemos luego ajustar un AR(3).

Vamos a aceptar el modelo propuesto si:

1. Los valores estimados de a_3 no resulta significativamente diferente de cero, y
2. Los valores estimados de a_1, a_2 no varían mucho respecto de las estimaciones originales.

2. Overfitting y redundancia de parámetros

Cuando tenemos un modelo ARMA, se presenta el problema de redundancia de parámetros o falta de identificabilidad. Si $a(B)Y_t = b(B)e_t$ es el modelo correcto, luego también es correcto el modelo $(1 - cB)a(B)Y_t = (1 - cB)b(B)e_t$ para cualquier constante *c*. Sin embargo, si el modelo original se correspondía con un ARMA(p,q), el segundo es un ARMA(p+1,q+1). Decimos en este caso que hay redundancia de parámetros.

1. Especificamos el modelo más sencillo que se vea factible (antes de probar alguno más complejo)
2. Al hacer overffiting, agrandamos la parte MA y la AR por separado
3. Expandir el modelo en la dirección sugerida por el análisis de residuos.

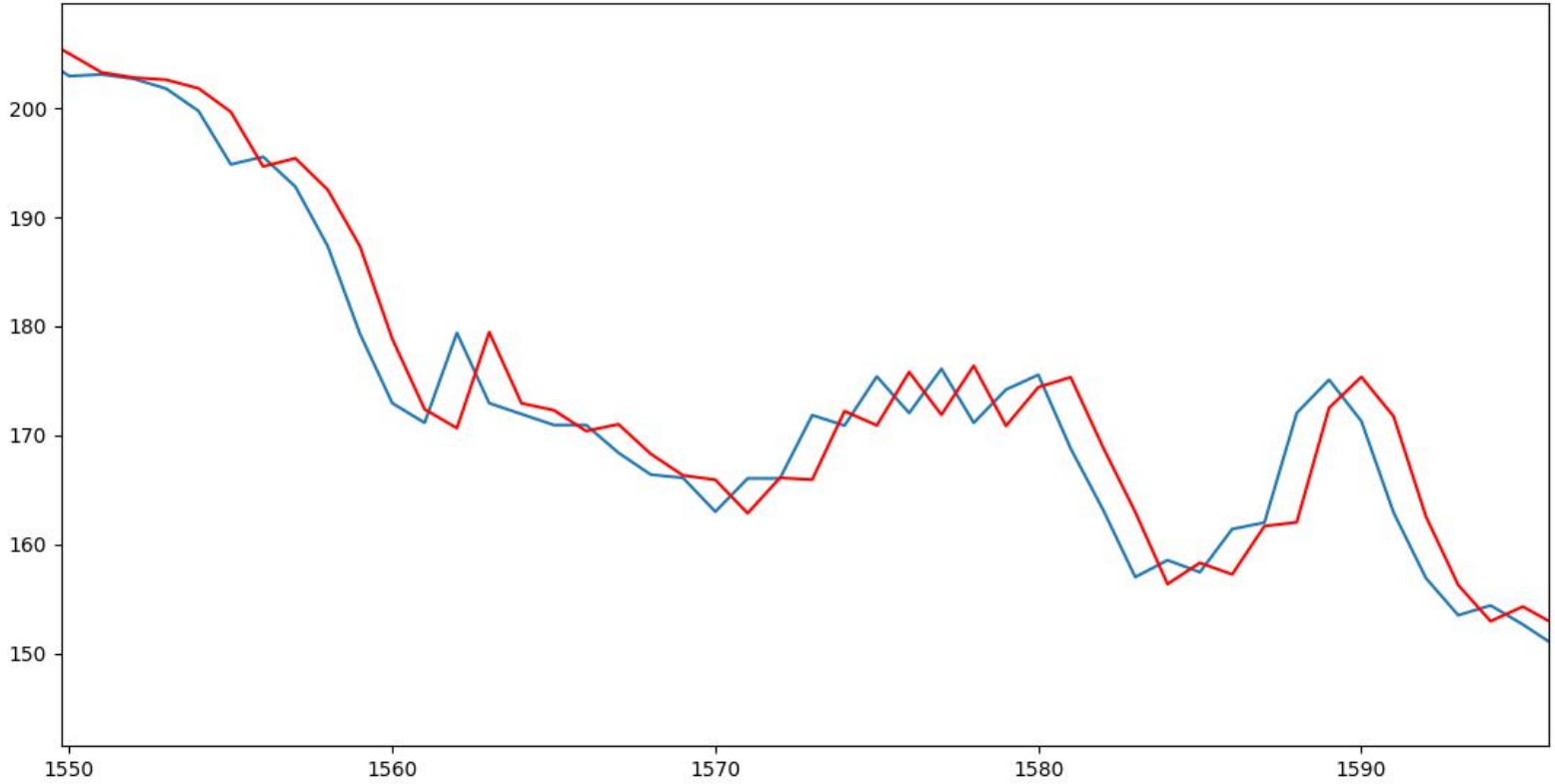
Ejemplo de ajuste usando ARIMA

```
# fit model
model = ARIMA(ts.ultimoPrecio.values, order=(5,1,0))
model_fit = model.fit()
# summary of fit model
print(model_fit.summary())
# line plot of residuals
residuals = DataFrame(model_fit.resid)
residuals.plot()
pyplot.show()
# density plot of residuals
residuals.plot(kind='kde')
pyplot.show()
# summary stats of residuals
print(residuals.describe())
```

Ejemplo de ajuste usando ARIMA

```
# evaluate an ARIMA model using a walk-forward validation
X = ts.ultimoPrecio.values
size = int(len(X) * 0.66)
train, test = X[0:size], X[size:len(X)]
history = [x for x in train]
predictions = list()
# walk-forward validation
for t in range(len(test)):
    model = ARIMA(history, order=(5,1,0))
    model_fit = model.fit()
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    #print('predicted=%f, expected=%f' % (yhat, obs))
```

Ejemplo de ajuste usando ARIMA



Resultados (S) ARIMA

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 4840
Model:                 ARIMA(5, 1, 0)   Log Likelihood              -10901.009
Date:                Thu, 11 Nov 2021   AIC                          21814.019
Time:                        17:51:37   BIC                          21852.926
Sample:                             0   HQIC                         21827.678
                                 - 4840
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0367      0.004      9.613      0.000       0.029       0.044
ar.L2          0.0218      0.005      4.612      0.000       0.013       0.031
ar.L3          0.0451      0.005      9.124      0.000       0.035       0.055
ar.L4         -0.0205      0.005     -3.941      0.000      -0.031      -0.010
ar.L5          0.0056      0.005      1.041      0.298      -0.005       0.016
sigma2         5.2995      0.027    199.035      0.000       5.247       5.352
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):            227927.17
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):             953.07   Skew:                            -0.30
Prob(H) (two-sided):                  0.00   Kurtosis:                        36.62
===================================================================================
```

Modelos estacionales

Modelo ARIMA estacional (SARIMA)

Tenemos modelos aditivos y modelos multiplicativos. Los más comunes son de este último tipo.

Modelo SARMA multiplicativo

Cuando el proceso es estacionario podemos definir un modelo SARMA multiplicativo **ARMA(p,q)x(P,Q)s**, con período estacional **s** como un modelo AR con polinomio característico $a(x)\alpha(x)$ más un MA con polinomio característico $b(x)\beta(x)$. Con

$$
\begin{aligned}
a(x) &= 1 - a_1 x - a_2 x^2 - \ldots - a_p x^p \\
\alpha(x) &= 1 - \alpha_1 x^s - \alpha_2 x^{2s} - \ldots - \alpha_P x^{Ps}, \\
b(x) &= 1 - b_1 x - b_2 x^2 - \ldots b_q x^q \\
\beta(x) &= 1 - \beta_1 x^s - \beta_2 x^{2s} - \ldots - \beta_P x^{Ps},
\end{aligned}
$$

Modelo SARIMA multiplicativo

Una herramienta importante es el análisis de procesos estacionales **no estacionarios** es la diferenciación estacional de período s para la serie {Yt}, denotada

$$\nabla_s Y_t = Y_t - Y_{t-s}$$

Se dice que una serie estacional no estacionaria sigue un modelo SARIMA(p,d,q)x(P,D,Q)s de período s si la serie diferenciada

$$W_t = \nabla^d \nabla_s^D Y_t$$

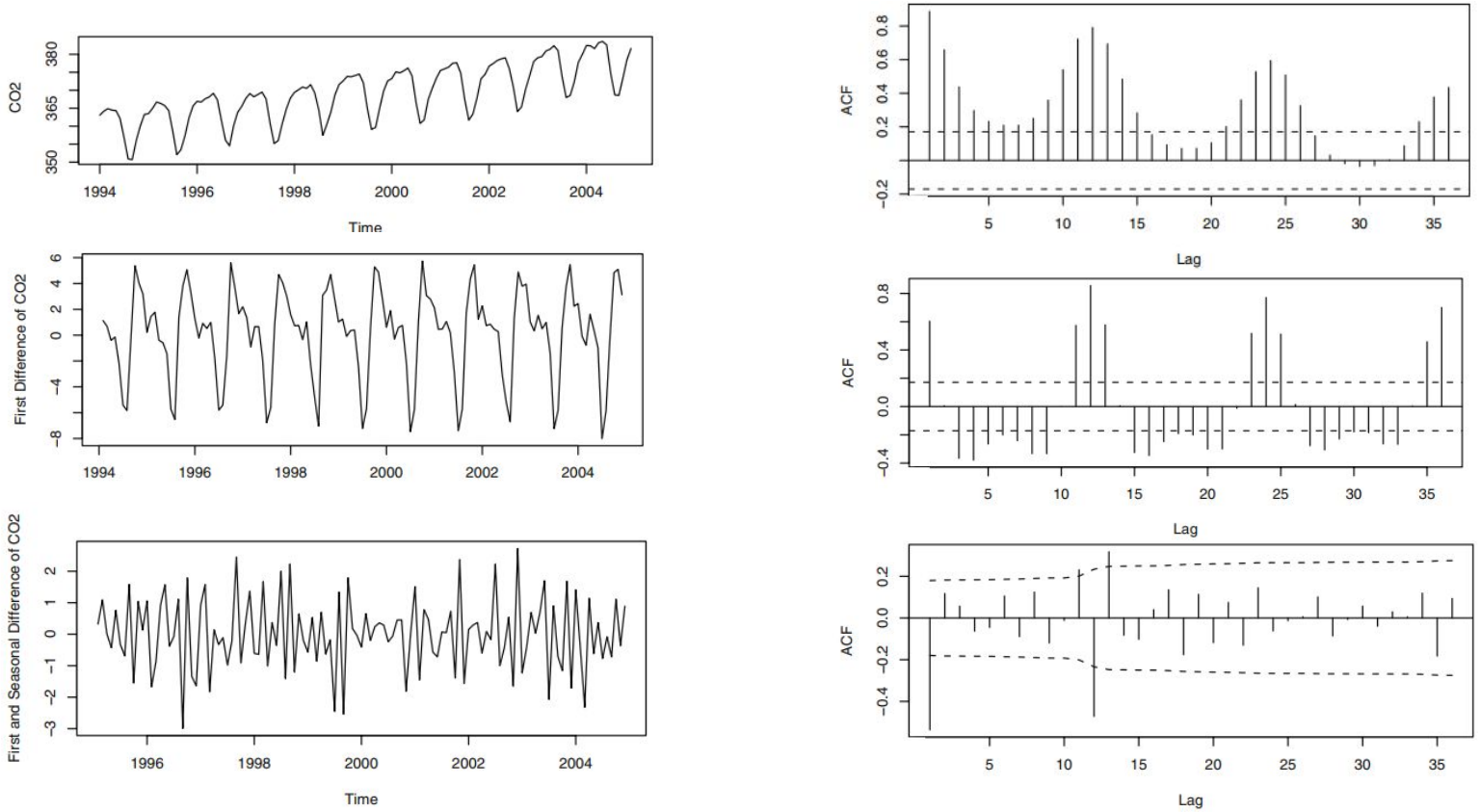
sigue un proceso SARMA(p,q)x(P,Q)s.

Especificación del modelo

Se usan las mismas ideas introducidas para modelos ARMA. y ARIMA.

1) Inspeccionar la serie de tiempo y su función de autocorrelación muestral. ¿Observo alguna tendencia y/o algún comportamiento estacional?
2) Proponer transformaciones (ej. diferenciar) y volver a analizar la serie resultante

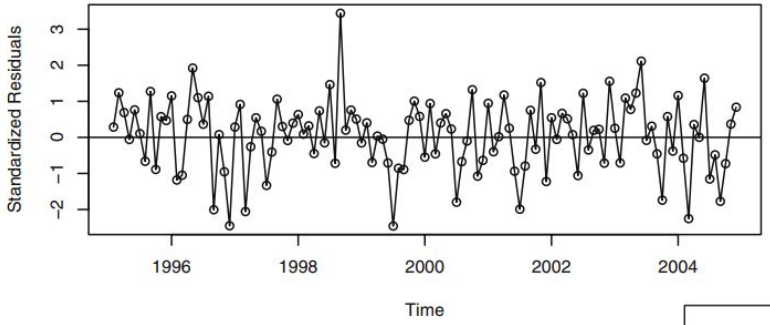
Especificación del modelo - ejemplo

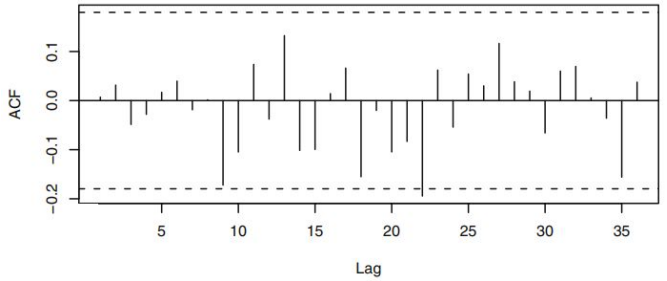


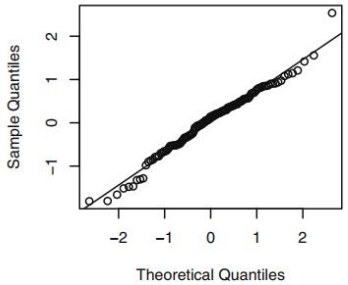
Ajuste del modelo

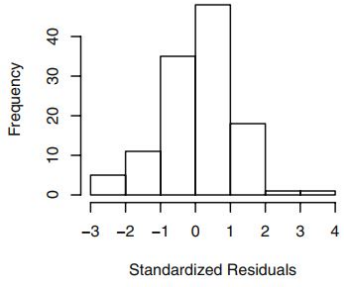
Una vez especificado el modelo (en el ejemplo anterior un candidato sería SARIMA(0,1,1)x(0,1,1)$_{12}$) debemos ajustar los parámetros.

Nuevamente vamos a analizar los residuos de la estimación



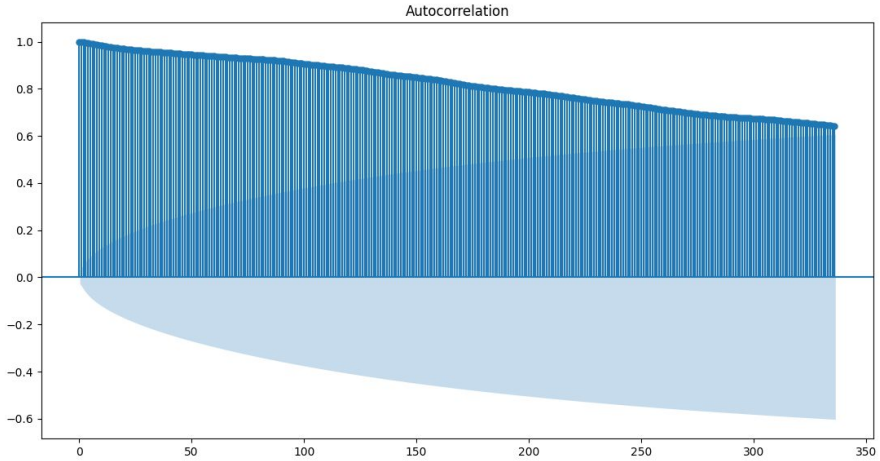






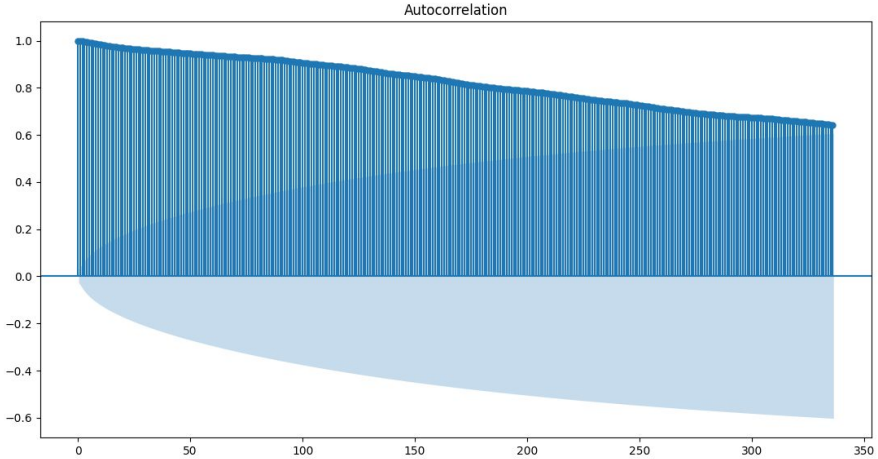
Función de Autocorrelación

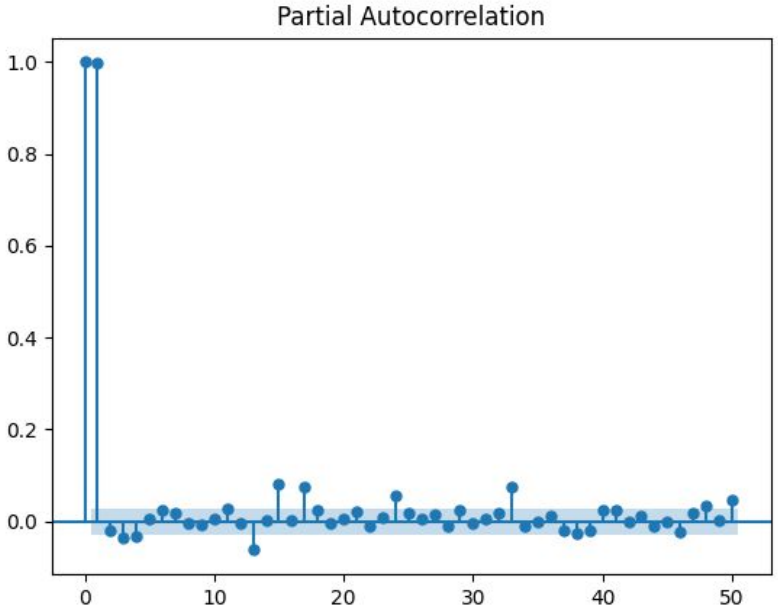
Función de autocorrelación



→ Indicio de que la serie es NO estacionaria

Función de autocorrelación parcial

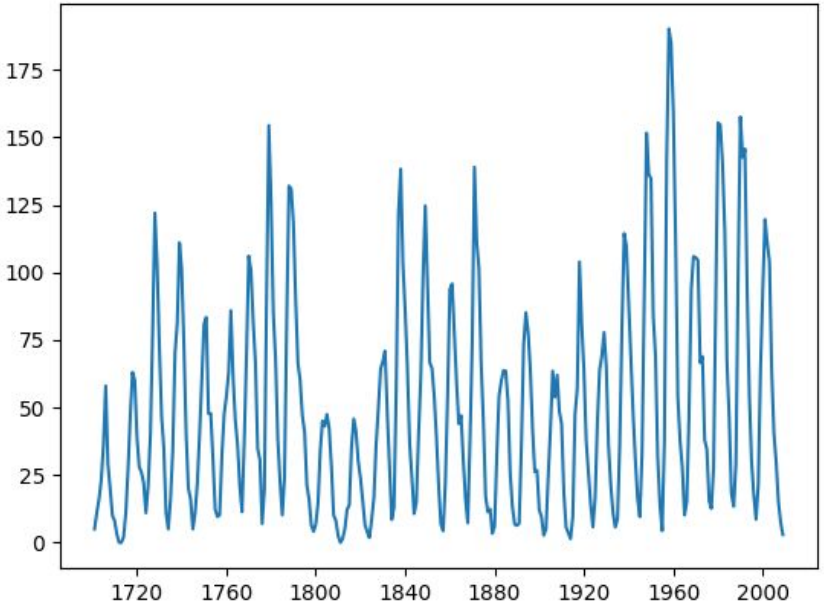




Función de autocorrelación

Ejemplo con dataset público

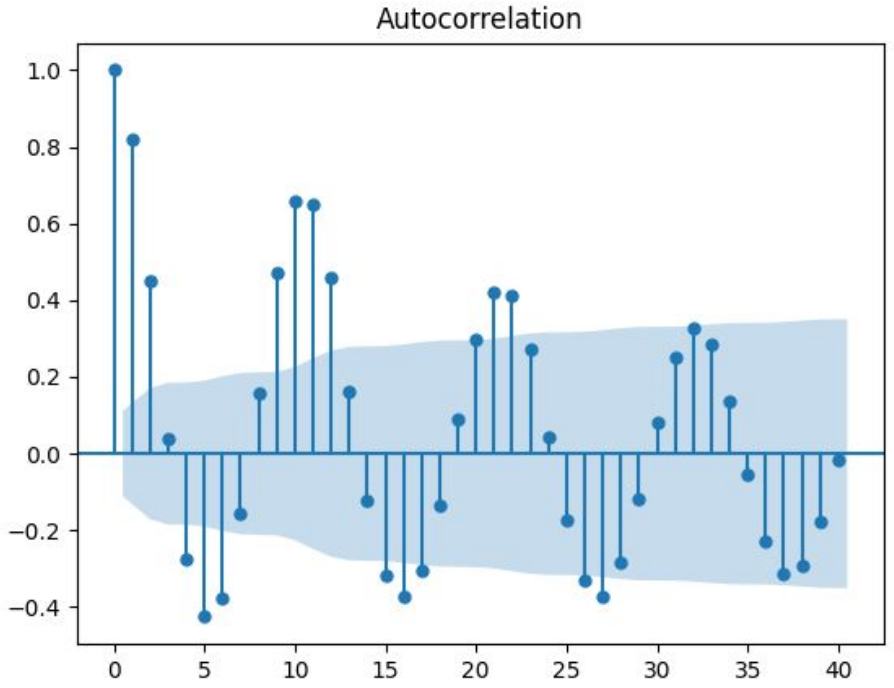
```
import statsmodels.api
as sm

dta =
sm.datasets.sunspots.loa
d_pandas().data
```



Función de autocorrelación

Ejemplo con dataset público

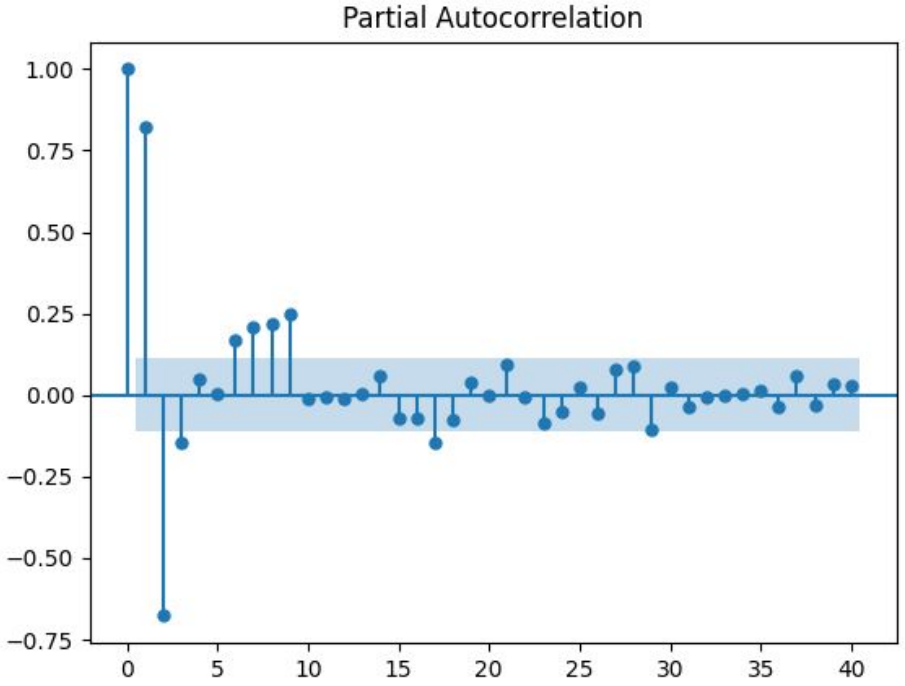
```
sm.graphics.tsa.plot_acf
(dta.values.squeeze(),
lags=40)
```



Función de autocorrelación parcial

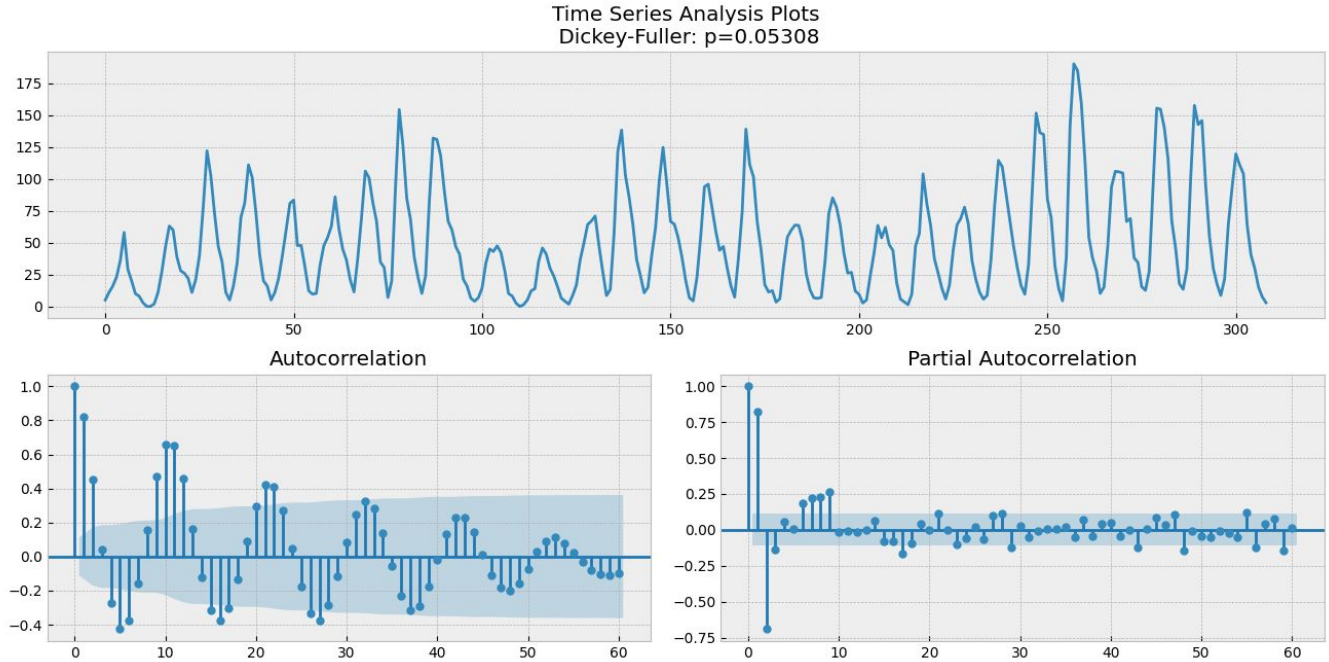
Ejemplo con dataset público

```
sm.graphics.tsa.plot_pacf
(dta.values.squeeze(),
lags=40, method="ywm")
```



Ajuste de modelo SARIMAX

Serie de actividad del sol (SUNACTIVITY)



Ajuste de modelo SARIMAX

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                          SUNACTIVITY   No. Observations:                  309
Model:             SARIMAX(2, 1, 3)x(1, 1, [1], 12)   Log Likelihood               -1251.921
Date:                              jue, 11 nov 2021   AIC                           2519.842
Time:                                      19:44:27   BIC                           2549.364
Sample:                                           0   HQIC                          2531.662
                                                - 309
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.6088      0.031     51.331      0.000       1.547       1.670
ar.L2         -0.9329      0.027    -34.808      0.000      -0.985      -0.880
ma.L1         -1.4445      0.063    -22.944      0.000      -1.568      -1.321
ma.L2          0.4701      0.104      4.539      0.000       0.267       0.673
ma.L3          0.1218      0.065      1.877      0.061      -0.005       0.249
ar.S.L12      -0.0102      0.075     -0.136      0.892      -0.158       0.138
ma.S.L12      -0.9980      1.843     -0.542      0.588      -4.610       2.614
sigma2       241.3306    439.808      0.549      0.583    -620.678    1103.339
==========================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                23.97
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               1.41   Skew:                             0.43
Prob(H) (two-sided):                  0.09   Kurtosis:                         4.10
==========================================================================================
```

Ajuste de modelo SARIMAX

```
best_model=sm.tsa.statespace. SARIMAX(ads.Ads, order= (p, d, q),
                                      seasonal_order= (P, D, Q, s)).fit(disp=-1)
```

el análisis de acf() y pacf() nos debe dar indicios de cómo elegir los parámetros (p,d,q)

Adicionalmente, el período **s** debe salir claro de las gráficas, en este caso es un período estacional igual a 12.