Modelos ARIMA

Gustavo A. García

ggarci24@eafit.edu.co

Econometría II

Programa de Economía

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Introducción
- El enfoque Box-Jenkins
- Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

Lecturas

- Maddala, GS. y Lahiri, K. (2009). *Introduction to econometrics*. 4a edición, Willey. Cap 13
- Enders, W. (2014). *Applied econometric time series*. 4th edition, Wiley. Cap 2, sección 8
- Pfaff, B. (2008). *Analysis integrated and cointegrated series with R*. 2th edition, Springer. Part I, sección 1.4
- Hyndman, R.J., y Athanasopoulos, G. (2021). *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia.

Páginas webs

- https://finnstats.com/index.php/2021/04/26/timeseries-analysis-in-r/

Introducción

- Hemos estudiado cómo una serie de tiempo puede ser explicada o bien por su historia (AR(p)) o por choques contemporáneos o pasados (MA(q))
- También estudiamos que estos dos procesos pueden ponerser juntos en un proceso más general ARMA(p,q)
- Ahora estudiaremos brevemente los modelos ARIMA o el enfoque de Box-Jenkins para series de tiempo
- Este enfoque consiste en tres etapas:
    1. identificación
    2. estimación
    3. diagnóstico

El enfoque Box-Jenkins

- El enfoque de Box-Jenkins (BJ) es una de las metodologías más amplio para el análisis de las series de tiempo

- Los pasos básicos de la metodología BJ son:
  1. diferenciar la serie, de modo que se alcance la estacionariedad
  2. identificar un modelo tentativo
  3. estimar el modelo
  4. verificar el diagnóstico (si se encuentra que el modelo es inadecuado, volver al paso 2)
  5. usar el modelo para pronosticar

El enfoque Box-Jenkins

Primer paso

- Determinar si la series es estacionaria
  - correlograma
  - test de raices unitarias
- Si la serie no es estacionaria deferenciarla (cuantas veces sea necesario) para logra su estacionariedad

Segundo paso

- Examinar el correlograma de la serie estacionaria para decidir los ordenes apropiados de los componentes AR y MA

Tercer paso

- Estimación del modelo ARMA

Cuarto paso

- Verificación del diagnóstico para comprobar la ideonidad del modelo tentativo

Quinto paso

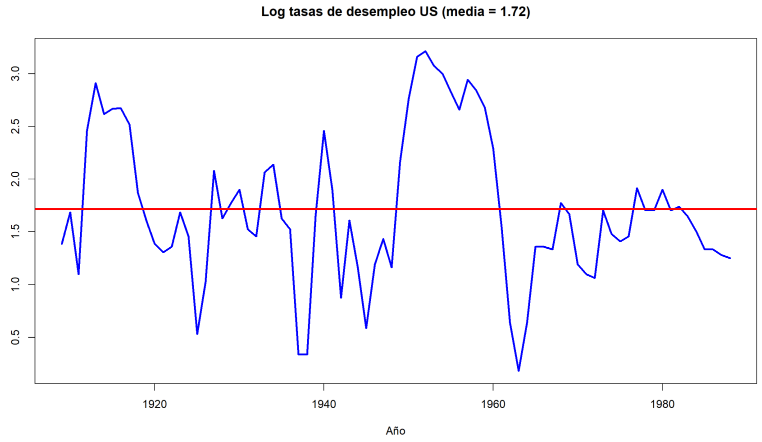
- Realizar el pronóstico con el modelo ARIMA

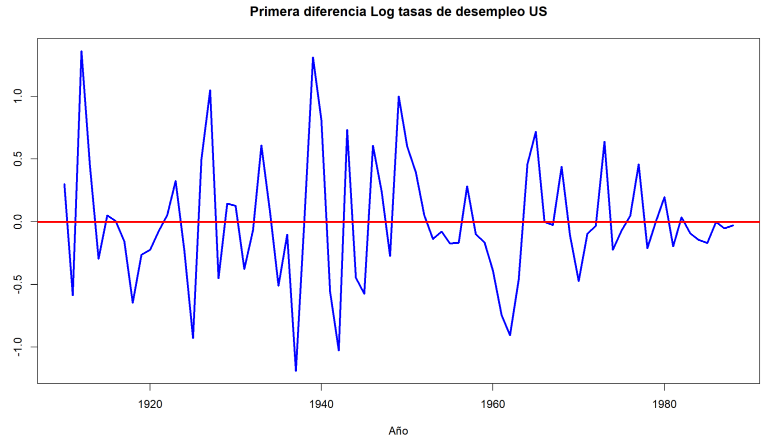
Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

```
library(urca); library(tidyverse); library(forecast); library(stats); library(lmtest)
data(npext)

y <- ts(na.omit(npext$unemploy), start=1909, end=1988, frequency = 1)
```

```
ts.plot(y, main = "Log tasas de desempleo US (media = 1.72)", xlab = "Año"
abline(h = mean(y), col = "red", lwd = 3)
```

```
ts.plot(diff(y), main = "Primera diferencia Log tasas de desempleo US", xl
abline(h = 0, col = "red", lwd = 3)
```





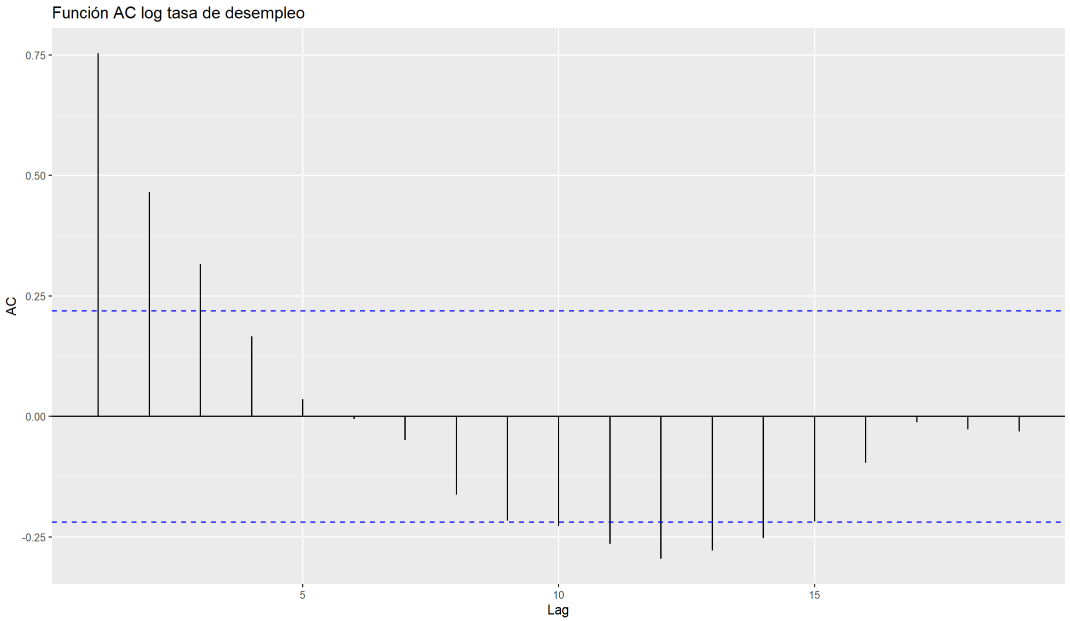
Parece preferible trabajar con la series sin diferenciar:

- La serie en log no tiene tendencia
- Existe buena cantidad de persistencia, en el sentido que las duraciones cuando el diferencial está por encima o por debajo de la media son algo largas
- La dinámica de la serie parece ser constante sobre la media
- Cuando se hace el Dickey-Fuller indica que la serie es estacionaria
- Parace que la serie es estacionaria en covarianza
- La serie diferenciada es muy errática y podría tener poco contenido informativo para hacer predicción de valores futuros

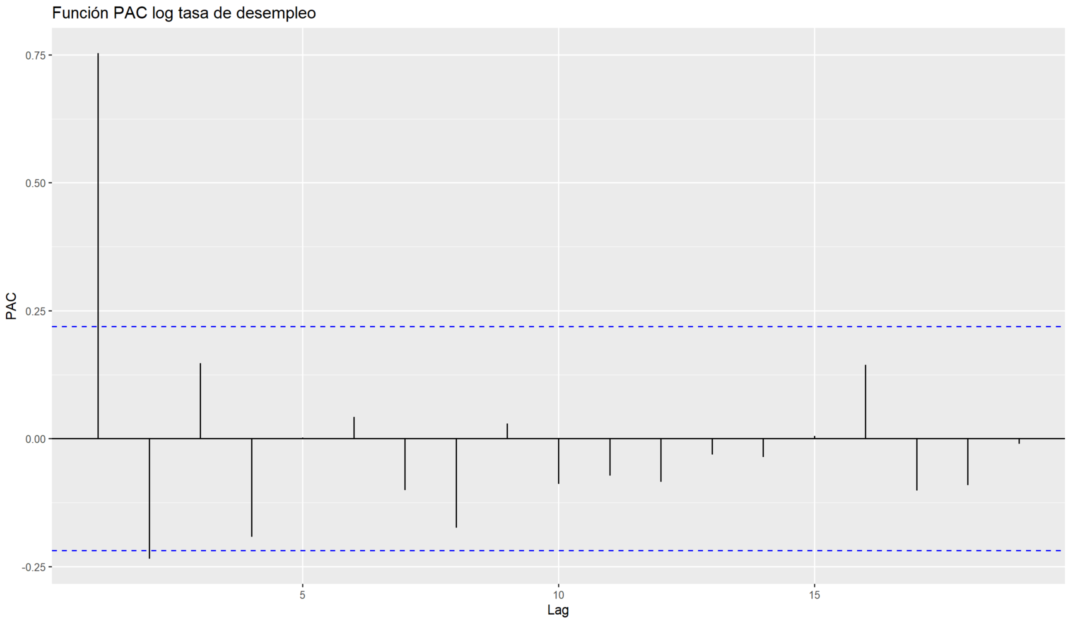
Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

Ahora calculamos el correlograma de la serie

```
ggAcf(y, main="Función AC log tasa de desempleo", ylab="AC")
```




```
ggPacf(y, main="Función PAC log tasa de desempleo", ylab="PAC")
```



Parece ser un proceso ARMA(2,0). Estimemos también un proceso sobreparametrizado ARMA(3,0) y pongamos a competir los dos procesos

Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

```
arma20 <- arima(y, order=c(2,0,0))
arma20
```

```
Call:
arima(x = y, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      0.9297  -0.2356     1.6988
s.e.  0.1079   0.1077     0.1586

sigma^2 estimated as 0.195:  log likelihood = -48.59,  aic = 105.18
```

```
coeftest(arma20)
```

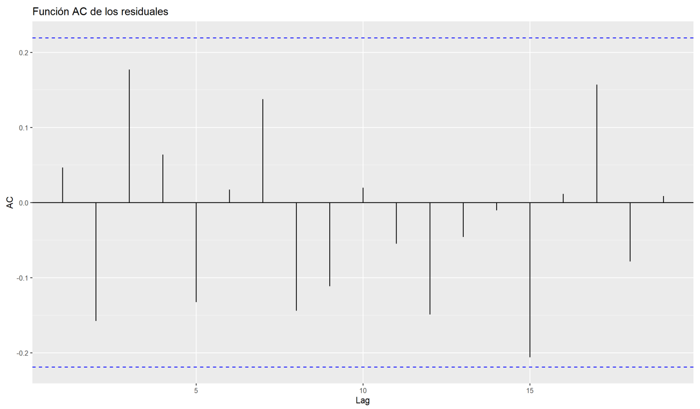
```
z test of coefficients:

          Estimate Std. Error z value Pr(>|z|)
ar1        0.92970    0.10786  8.6197  < 2e-16 ***
ar2       -0.23560    0.10771 -2.1874  0.02872 *
intercept  1.69883    0.15860 10.7116  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los parámetros estimados son estadísticamente significativos y los valores satisfacen la condición de estabilidad $(|\rho| < 1)$

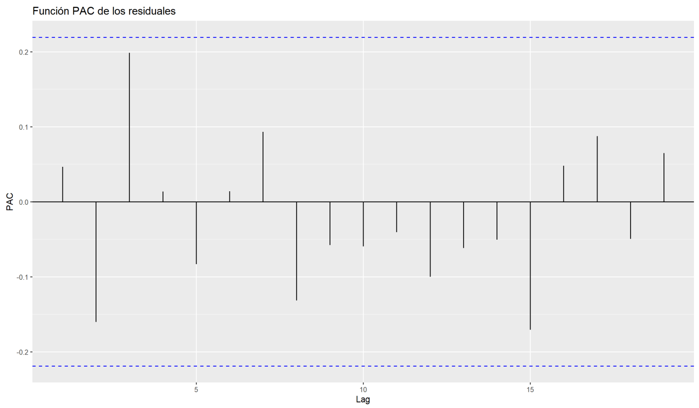
Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

Ahora pasamos a chequear los residuales del modelo

```
res20 <- residuals(arma20)
ggAcf(res20, main="Función AC de los residuales", ylab="AC")
```




```
ggPacf(res20, main="Función PAC de los residuales", ylab="PAC")
```




```
Box.test(res20, lag = 1, type = "Ljung-Box")
```

```
    Box-Ljung test

data:  res20
X-squared = 0.1797, df = 1, p-value = 0.6716
```

```
shapiro.test(res20)
```

```
    Shapiro-Wilk normality test

data:  res20
W = 0.99313, p-value = 0.9501
```

Se tiene que los residuales no están correlacionados y se distribuyen normal

Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

Comparemos con el modelo sobreparametrizado ARMA(3,0)

```
arma30 <- arima(y, order=c(3,0,0))
arma30
```

```
Call:
arima(x = y, order = c(3, 0, 0))

Coefficients:
         ar1      ar2     ar3  intercept
      0.9727  -0.3949  0.1669     1.6863
s.e.  0.1101   0.1495  0.1103     0.1851

sigma^2 estimated as 0.1893:  log likelihood = -47.47,  aic = 104.93
```

```
coeftest(arma30)
```

```
z test of coefficients:

          Estimate Std. Error z value  Pr(>|z|)    
ar1        0.97271    0.11013  8.8322 < 2.2e-16 ***
ar2       -0.39486    0.14954 -2.6405  0.008279 ** 
ar3        0.16690    0.11031  1.5130  0.130286    
intercept  1.68631    0.18511  9.1097 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Se observa que los coeficientes del primer y segundo rezago son similares al modelo ARMA(2,0), pero el tercer rezago no es estadísticamente diferente de cero
- Comparando el AIC (el mejor modelo es que tenga menor valor), ARMA(3,0) es mejor que ARMA(2,0)
- Un test likelihood-ratio puede calcularse para seleccionar el mejor modelo $(H_0:$ se prefiere el modelo restringido (ARMA(2,0)))

```
lrtest <- as.numeric(2*(logLik(arma30)-logLik(arma20)))
pchisq(lrtest, df=1, lower.tail = F)
```

```
[1] 0.1336066
```

Esto indica no rechazar H_0, es decir que las mejoras en el log-likelihood no son significantes de pasar de un modelo ARMA(2,0) a ARMA(3,0), por lo que se prefiere el modelo más parsimonioso ARMA(2,0)

Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

Una vez se ha estimado un modelo ARMA, puede ser usado para predecir valores futuros de la variable de interes

Estas predicciones pueden ser calculadas recursivamente desde el predictor lineal

$$\begin{aligned} Y_T(h) = & \rho_1 \overline{Y}_{T+h-1} + \ldots + \rho_1 \overline{Y}_{T+h-p} + \\ & \epsilon_t + \theta_1 \epsilon_{t-T-1} + \ldots + \epsilon_{t-T-q} \end{aligned}$$

donde $\overline{Y}_t = Y_t$ para $t \leq T$ y $\overline{Y}_{T+j} = Y_T(j)$ para $j = 1, \ldots, h-1$

Este predictor es equivalente a

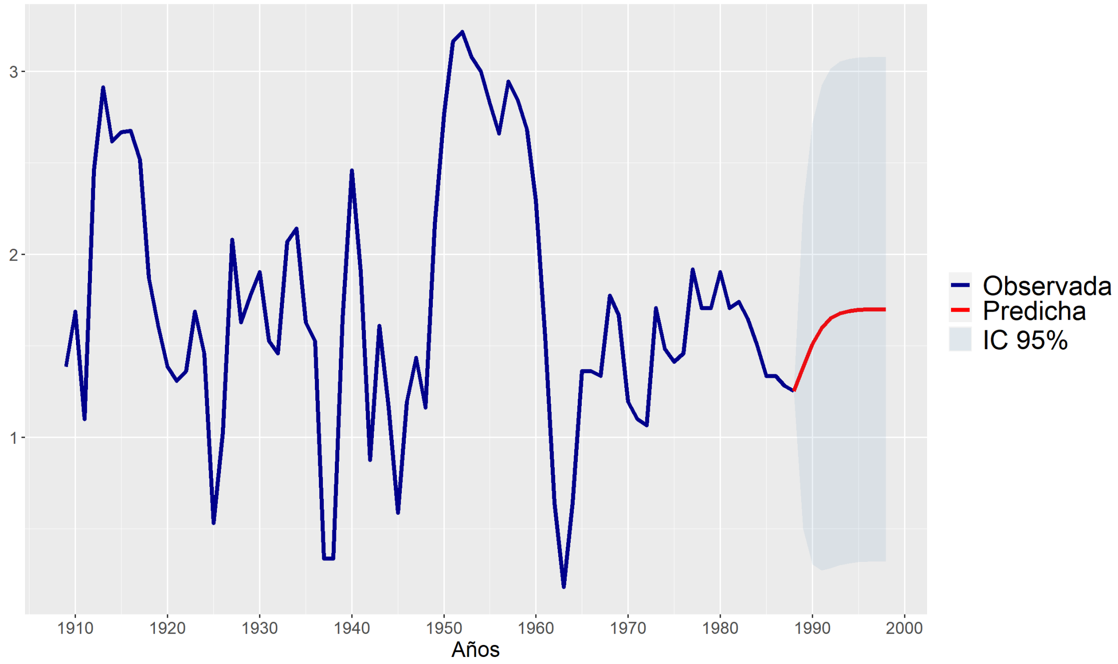
$$Y_T(h) = \mu + \psi_h \epsilon_t + \psi_{h+1} \epsilon_{t-1} + \psi_{h+2} \epsilon_{t-2} + \ldots$$

Cuando el horizonte de predicción h es mayor que el orden q del MA, la predicción son determinandas sólo por los términos autorregresivos

Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

Vamos a predecir 10 años de la tasa de desempleo

```
arma20.pred <- predict(arma20, n.ahead = 10)
predict <- ts(c(rep(NA, length(y) - 1), y[length(y)], arma20.pred$pred), start=1909, frequency = 1)
upper <- ts(c(rep(NA, length(y) - 1), y[length(y)], arma20.pred$pred + 2*arma20.pred$se), start = 1909, frequency = 1)
lower <- ts(c(rep(NA, length(y) - 1), y[length(y)], arma20.pred$pred-2*arma20.pred$se), start = 1909, frequency = 1)
observed <- ts(c(y, rep(NA, 10)), start = 1909, frequency = 1)

data <- data.frame(year = 1909:1998, actual = observed, predicho = predict, ic_l = lower, ic_u = upper)
ggplot(data) +
  geom_line(aes(x = year, y = actual, color = "Observada"), linewidth = 1.5) +
  geom_line(aes(x = year, y = predicho, color = "Predicha"), linewidth = 1.5) + geom_ribbon(aes(x = year, y = predicho, ymin = ic_l, ymax = ic_u, fill
  theme(legend.text = element_text(size = 20), text = element_text(size=16), legend.spacing.y = unit(-0.4, "cm"), legend.background=element_blank()) +
```



Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

Existe otra función más poderosa que selecciona el mejor modelo ARIMA

```
arma_op <- auto.arima(y)
arma_op
```

```
Series: y
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1     ma1    mean
      0.5272  0.5487  1.6934
s.e.  0.1221  0.1456  0.1546

sigma^2 = 0.1917:  log likelihood = -46.51
AIC=101.01   AICc=101.55   BIC=110.54
```

```
coeftest(arma_op)
```

```
z test of coefficients:

          Estimate Std. Error z value  Pr(>|z|)
ar1        0.52717    0.12213  4.3166 1.585e-05 ***
ma1        0.54866    0.14558  3.7687 0.0001641 ***
intercept  1.69340    0.15461 10.9526 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
f <- forecast(arma_op, 10, 95)
autoplot(f, main="Predicción de la tasa de desempleo")
```