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
- Modelos ARIMA estacionales
- Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos
- Ejercicio aplicado en R: desempleo estacional 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

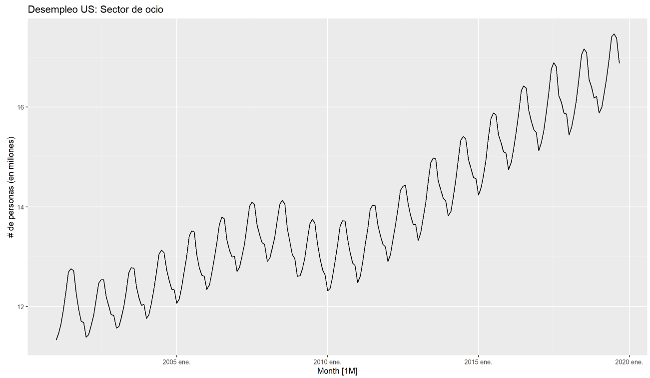
Quinto paso

- Realizar el pronóstico con el modelo ARIMA

Modelos ARIMA estacionales

Hasta ahora, hemos restringido la atención a datos no estacionales y modelo ARIMA no estacionales. Sin embargo, los modelos ARIMA también son capaces de modelar un amplio rango de dato estacionales. Los datos estacionales tienen la forma:

```r
library(fpp3)

data(us_employment)

leisure <- us_employment |>  filter(Title == "Leisure and Hospitality", year(Month) > 2000) |>
  mutate(Employed = Employed/1000) |> select(Month, Employed)
autoplot(leisure, Employed) + labs(title = "Desempleo US: Sector de ocio", y="# de personas (en millones)")
```



Un modelo estacional ARIMA incluye términos estacionales adicionales y tiene la siguiente estructura

$$ARIMA \ \underbrace{(p,d,q)}_{\text{No estacional}} \ \underbrace{(P,D,Q)_m}_{\text{Estacional}}$$

donde $m =$ el periodo estacional (ejemplo, número de observaciones por año)

Modelos ARIMA estacionales

Para definir la estructura estacional de una serie de tiempo se sigue el mismo procedimiento que con la parte nos estacional, es decir viendo el correlograma

Correlogragra de un proceso $ARIMA(0, 0, 0)(0, 0, 1)_{12}$

- AC: un pico estadísticamente significativo en el rezago 12 pero no significancia de otros picos
- PAC: decae exponencialmente en los rezagos estacionales (es decir, en los rezagos 12, 24, 36,...)

Correlogragra de un proceso $ARIMA(0, 0, 0)(1, 0, 0)_{12}$

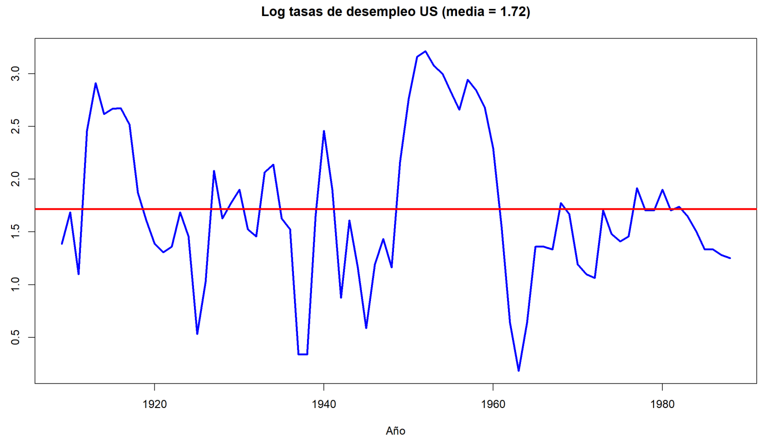
- AC: decae exponencialmente en los rezagos estacionales (es decir, en los rezagos 12, 24, 36,...)
- PAC: un pico estadísticamente significativo en el rezago 12 pero no significancia de otros picos

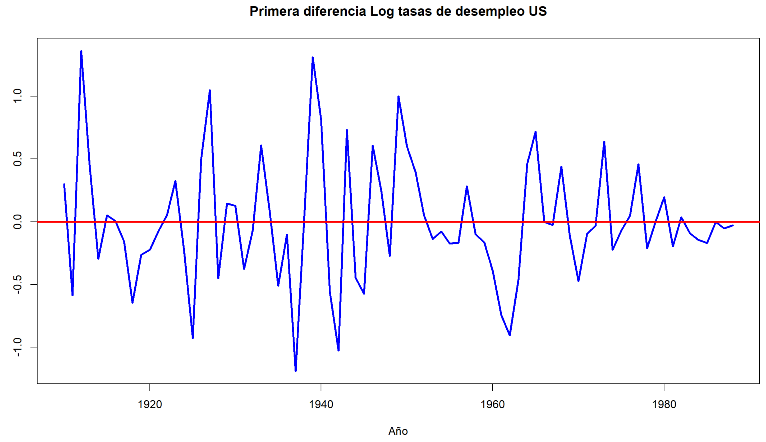
Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

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

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

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

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





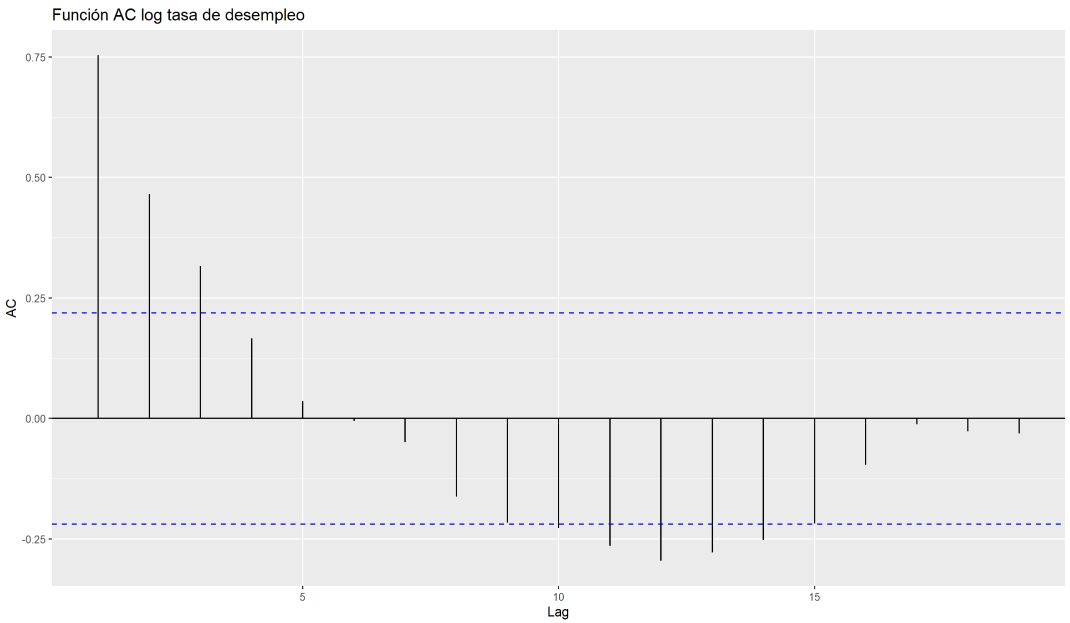
Parece preferible trabajar con la series sin diferenciar:

- 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

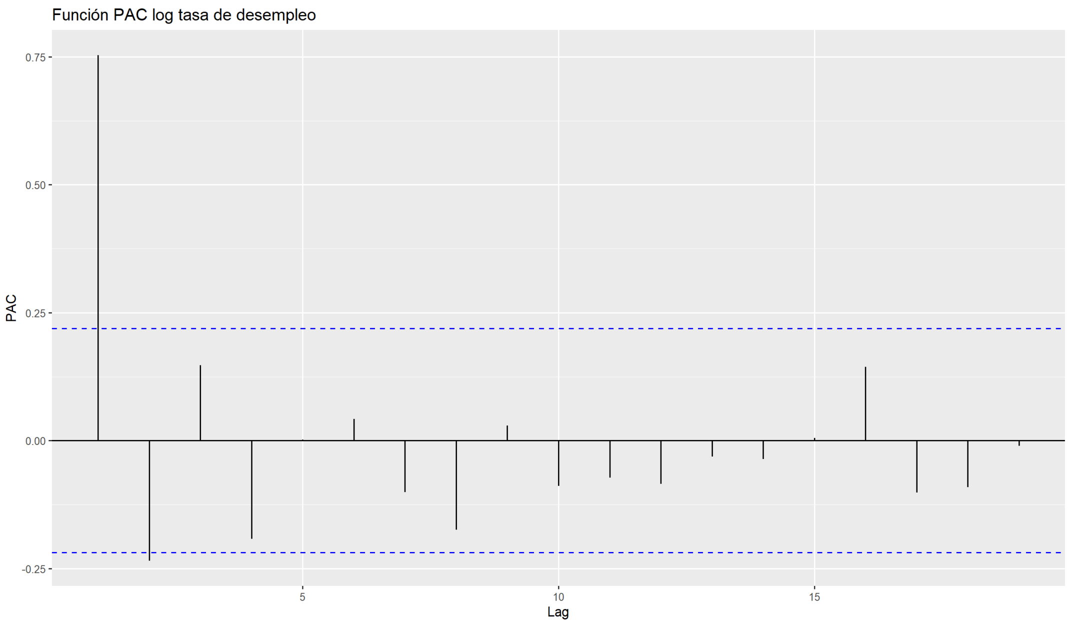
Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

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)
```

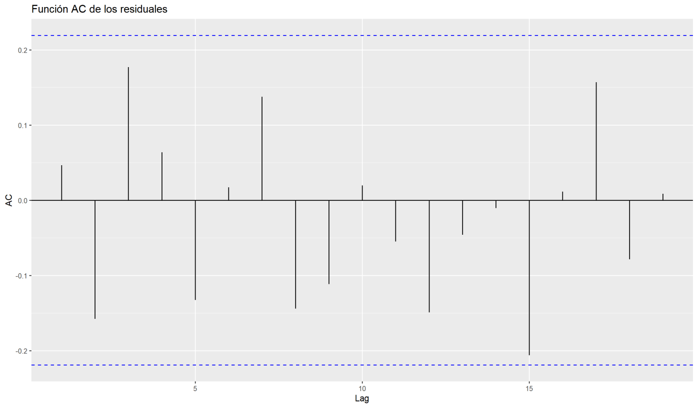
```
z test of coefficients:

          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 $(|\rho| < 1)$

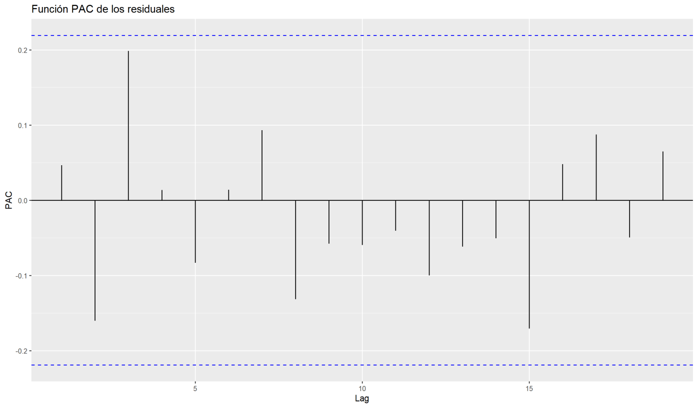
Ejercicio aplicado en R: tasa de desempleo de los Estados Unidos

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
```

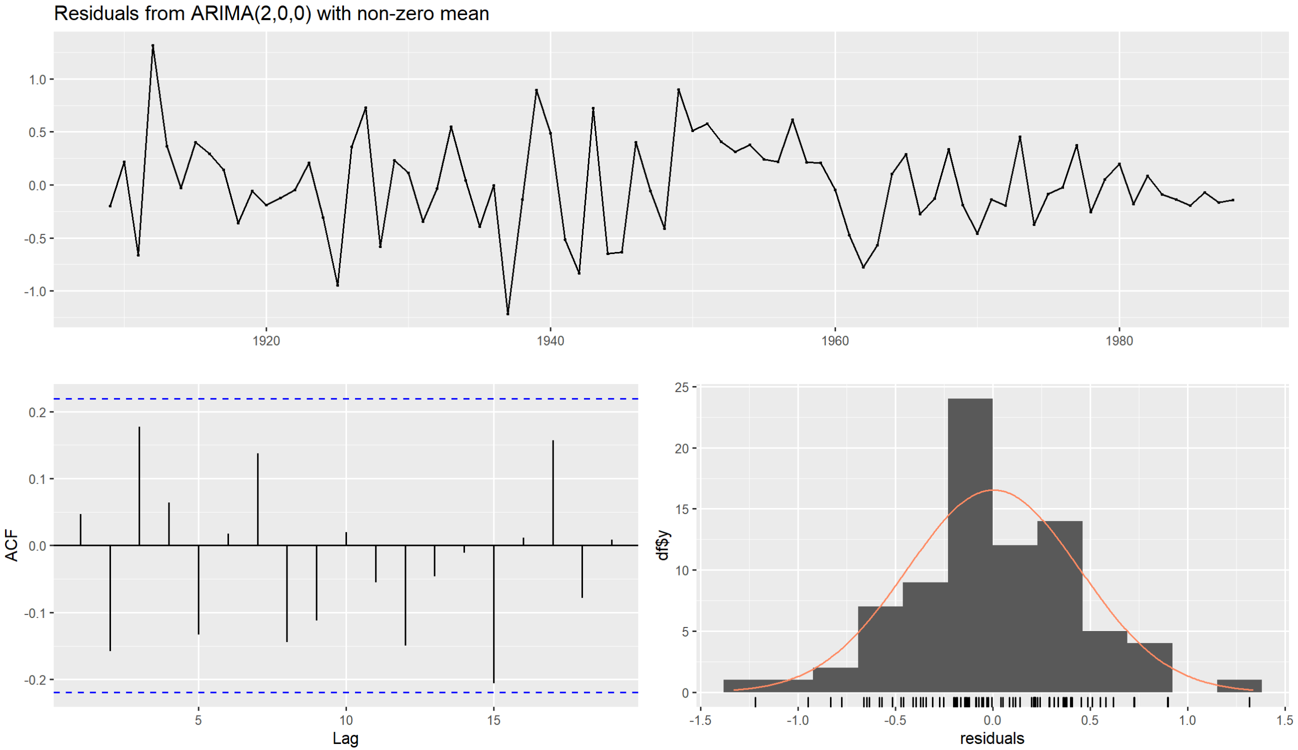
```
shapiro.test(res20)
```

```
    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

```
checkresiduals(arma20)
```




```
    Ljung-Box test

data:  Residuals from ARIMA(2,0,0) with non-zero mean
Q* = 11.648, df = 8, p-value = 0.1676

Model df: 2.   Total lags used: 10
```

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 los criterios de información (el mejor modelo es que tenga menor valor)

```
AIC(arma20); AIC(arma30)
```

```
[1] 105.1803
```

```
[1] 104.9302
```

```
BIC(arma20); BIC(arma30)
```

```
[1] 114.7084
```

```
[1] 116.8403
```

Con el AIC el ARMA(3,0) es mejor que el ARMA(2,0), pero con el BIC ARMA(2,0) es mejor que el ARMA(3,0)

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

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

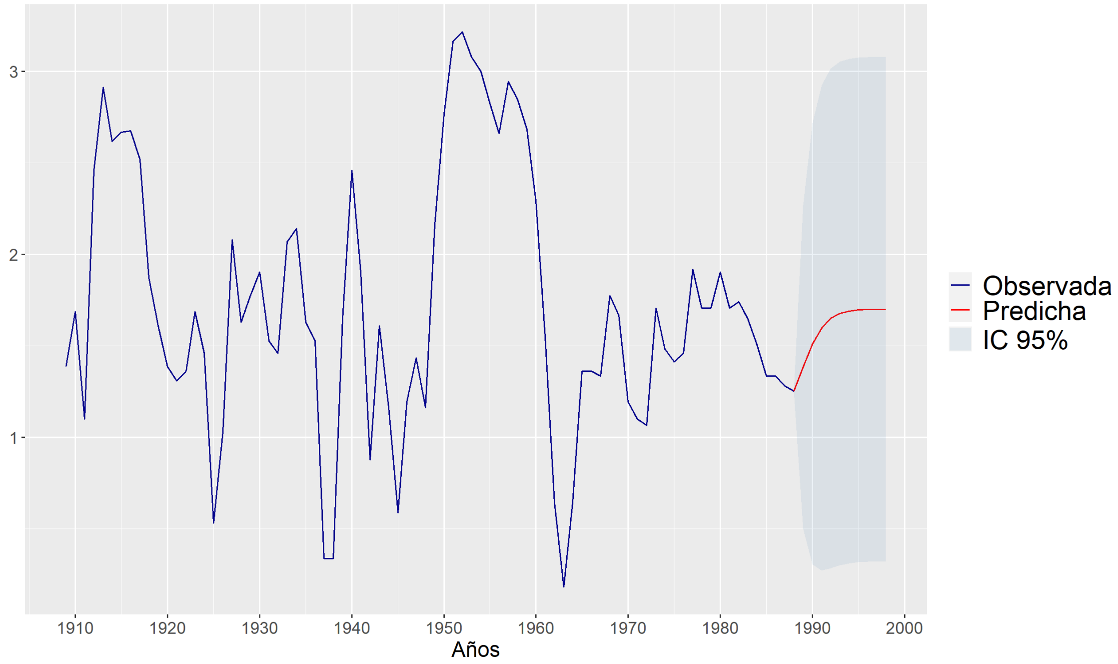
$$
Y_T(h) = \mu + \psi_h \epsilon_t + \psi_{h+1} \epsilon_{t-1} + \psi_{h+2} \epsilon_{t-2} + \ldots
$$

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

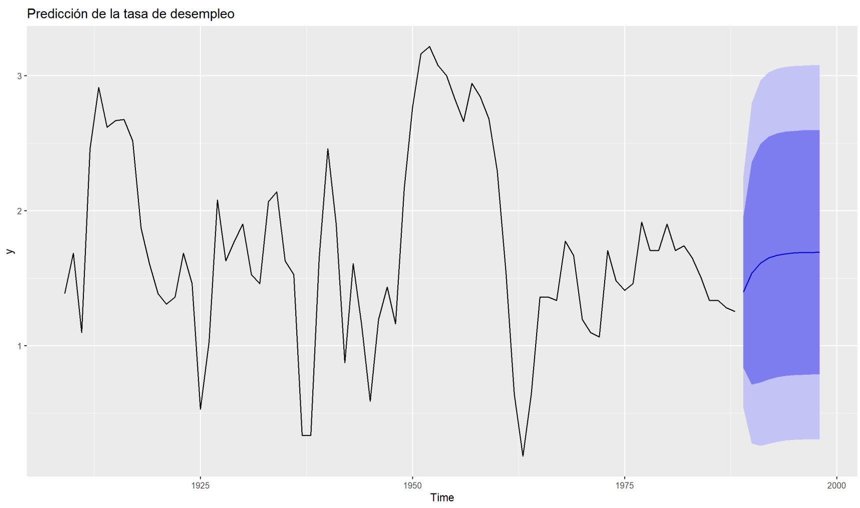
```
arma_op <- auto.arima(y, stepwise=F, approximation=F)
arma_op
```

```
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
```

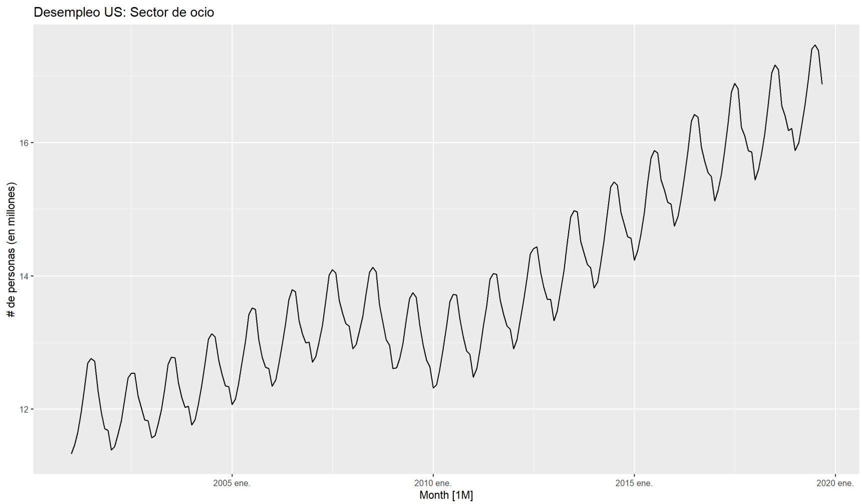
```
f <- forecast::forecast(arma_op, 10)
autoplot(f, main="Predicción de la tasa de desempleo")
```



Ejercicio aplicado en R: desempleo estacional en los Estados Unidos

En este ejemplo se decribe la modelación de un ARIMA estacional usando los datos mensuales de desempleo para Estados Unidos. Se toman los datos para el sector de ocio y hostelería desde enero de 2001 a septiembre de 2019

```r
data("us_employment")

leisure <- us_employment |>  filter(Title == "Leisure and Hospitality", year(Month) > 2000) |>
  mutate(Employed = Employed/1000) |> select(Month, Employed)
autoplot(leisure, Employed) + labs(title = "Desempleo US: Sector de ocio", y="# de personas (en millones)")
```

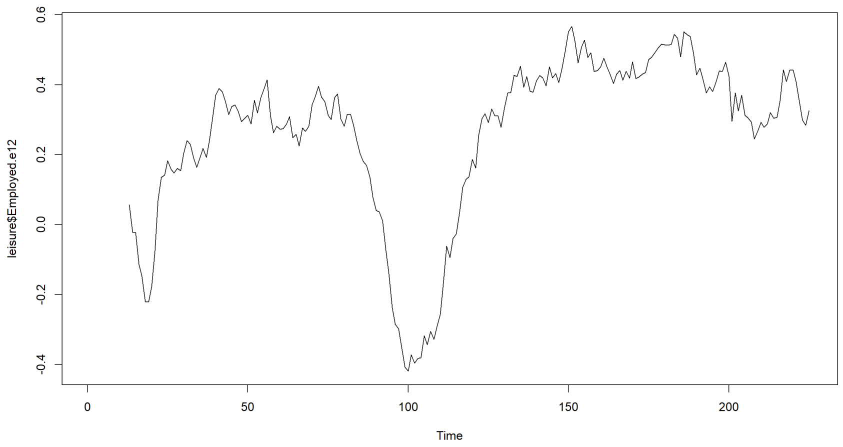


Los datos son claramente no estacionarios, con una fuerte estacionalidad y tendencia no lineal

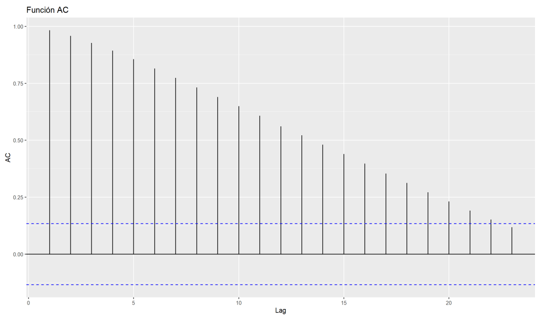
Ejercicio aplicado en R: desempleo estacional en los Estados Unidos

Se toma una diferencia estacional en 12 meses, para eliminar esa estacionalidad

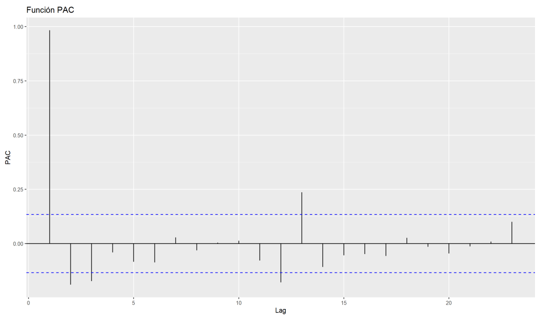
```
leisure <- leisure |> mutate(Employed.e12 = difference(leisure$Employed, 12))
ts.plot(leisure$Employed.e12)
```




```
ggAcf(leisure$Employed.e12, main="Función AC", ylab="AC")
```



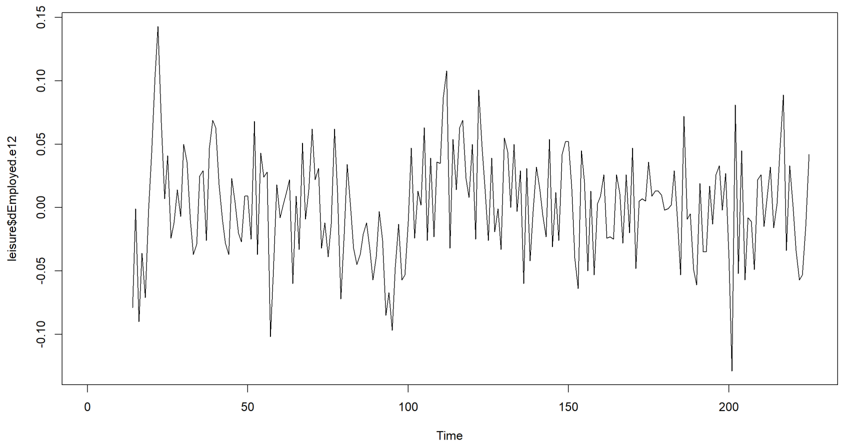

```
ggPacf(leisure$Employed.e12, main="Función PAC", ylab="PAC")
```



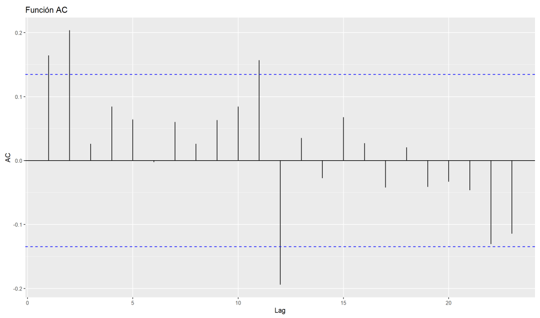
Ejercicio aplicado en R: desempleo estacional en los Estados Unidos

Se observa que no hay estacionariedad, así que se toman primeras diferencias y se vuelve a mirar el correlograma

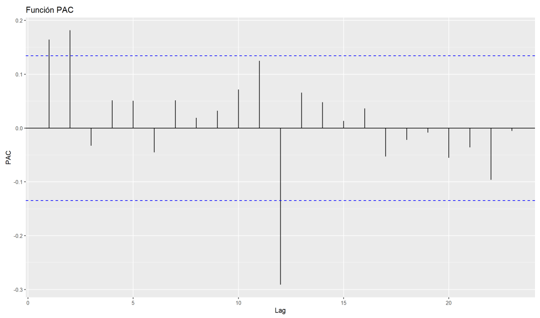
```r
leisure <- leisure |> mutate(dEmployed.e12 = difference(leisure$Employed.e12, 1))
ts.plot(leisure$dEmployed.e12)
```




```r
ggAcf(leisure$dEmployed.e12, main="Función AC", ylab="AC")
```




```r
ggPacf(leisure$dEmployed.e12, main="Función PAC", ylab="PAC")
```



Ejercicio aplicado en R: desempleo estacional en los Estados Unidos

Potenciales modelos que surgen con base en el anterior correlograma:

- la significancia en el rezago 2 de la AC sugiere un MA(2) en la parte no estacional. La significancia en el rezago 12 en la AC sugiere un MA(1) en la parte estacional. El proceso sería un ARIMA(0,1,2)(0,1,1)
- si se usa la PAC para seleccionar la parte no estacional y la AC para seleccionar la parte estacional, surge un ARIMA(2,1,2)(0,1,1)
- También se incluye la selección atomática

```
arima012011 <- arima(leisure$Employed, order=c(0,1,2),
                     seasonal = list(order = c(0, 1, 1), period = 12))
arima012011
```

```
Call:
arima(x = leisure$Employed, order = c(0, 1, 2), seasonal = list(order = c(0,
    1, 1), period = 12))

Coefficients:
         ma1     ma2     sma1
      0.2315  0.2167  -0.5006
s.e.  0.0707  0.0621   0.0814

sigma^2 estimated as 0.001433:  log likelihood = 391.45,  aic = -774.9
```

```
arima210011 <- arima(leisure$Employed, order=c(2,1,0),
                     seasonal = list(order = c(0, 1, 1), period = 12), met
arima210011
```

```
Call:
arima(x = leisure$Employed, order = c(2, 1, 0), seasonal = list(order = c(0,
    1, 1), period = 12), method = "ML")

Coefficients:
         ar1     ar2     sma1
      0.2102  0.1941  -0.4969
s.e.  0.0683  0.0679   0.0788

sigma^2 estimated as 0.001425:  log likelihood = 392.09,  aic = -776.19
```

```
autoarima <- leisure |>
  model(ARIMA(Employed, stepwise = FALSE, approx = FALSE))
report(autoarima)
```

```
Series: Employed
Model: ARIMA(2,1,0)(1,1,1)[12]

Coefficients:
         ar1     ar2    sar1     sma1
      0.1786  0.1855  0.3295  -0.7507
s.e.  0.0695  0.0679  0.1273   0.0936

sigma^2 estimated as 0.001415:  log likelihood=394.96
AIC=-779.92   AICc=-779.63   BIC=-763.14
```

Ejercicio aplicado en R: desempleo estacional en los Estados Unidos

```
AIC(arima012011); AIC(arima210011)
```

```
[1] -774.8994
```

```
[1] -776.189
```

```
BIC(arima012011); BIC(arima210011)
```

```
[1] -761.4731
```

```
[1] -762.7627
```

```
glance(autoarima)
```

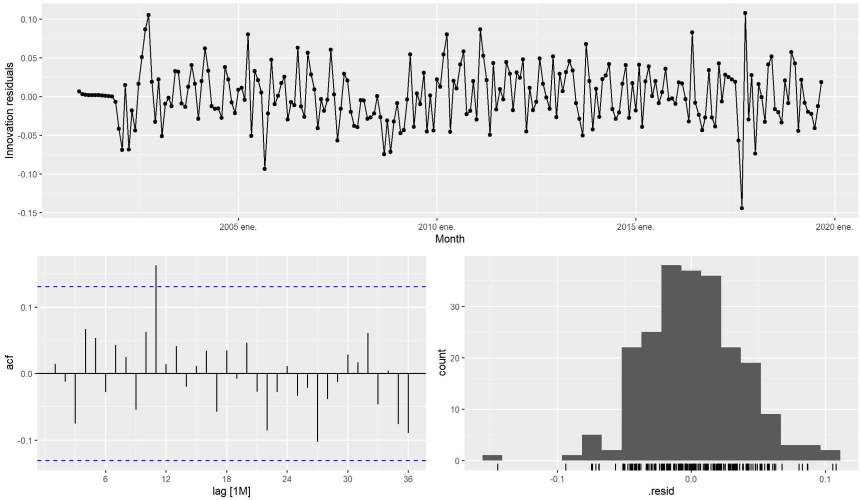
```
# A tibble: 1 × 8
  .model              sigma2 log_lik   AIC  AICc   BIC ar_ro…¹ ma_ro…²
  <chr>                <dbl>   <dbl> <dbl> <dbl> <dbl> <list>  <list>
1 ARIMA(Employed, s… 0.00142    395. -780. -780. -763. <cpl>   <cpl>
# … with abbreviated variable names ¹ar_roots, ²ma_roots
```

De acuerdo con los criterios de información, el mejor modelo es el de la selección automática

Ejercicio aplicado en R: desempleo estacional en los Estados Unidos

Miremos los residuales del modelo

```
autoarima |> gg_tsresiduals(lag=36)
```



Un pequeño pero significante pico (en el rezago 11) de 36 es aún consistente con un comportamiento ruido blanco. Para estar seguros, se calcula el test Ljung-Box

```
augment(autoarima) |>
  features(.innov, ljung_box, lag=24, dof=4)
```

```
# A tibble: 1 × 3
  .model                                             lb_stat lb_pvalue
  <chr>                                                <dbl>     <dbl>
1 ARIMA(Employed, stepwise = FALSE, approx = FALSE)     16.6     0.680
```

El pvalor es muy grande con lo que se confirma que los residuales son similares a un ruido blanco

Ejercicio aplicado en R: desempleo estacional en los Estados Unidos

Haciendo la predicción con el modelo seleccionado

```
fabletools::forecast(autoarima, h=36) |>
  autoplot(leisure) +
  labs(title = "Predicción del desempleo en el sector de ocio US",
       y="# de personas (en millones)")
```