Series de Tiempo 2018

Maestría en Estadística Aplicada, UNR
Unidad 8

Luis Damiano
damiano.luis@gmail.com

2018-05-04

Contenido

- Diagnóstico de residuos
  - Visualizaciones
  - Pruebas de hipótesis
- Criterios de selección de modelos
- Ejercicio: Ventas en supermercados
- Ejercicio: Producción de automóviles

Diagnóstico

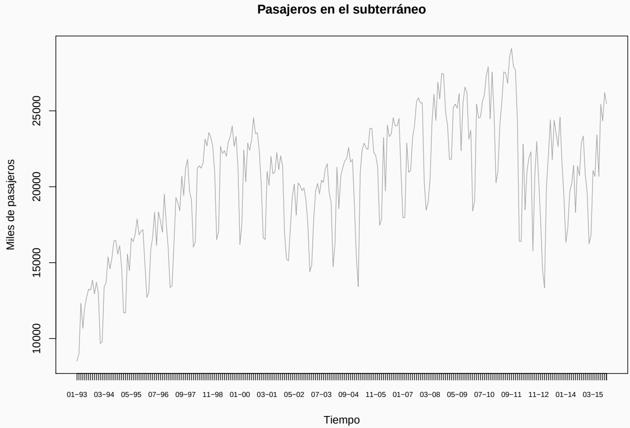
Supuestos

Supuestos:

$a_t \sim \mathrm{WN}(0, \sigma^2)$

- Distribución Gaussiana.
  - Histograma de residuos.
  - Gráfico de Cuantil-Cuantil de residuos.
  - Prueba de bondad de ajuste χ^2.
  - Pruebas de normalidad.
- Centrado en cero.
- Varianza constante.
  - Gráfico de residuos.
- Independencia condicional.
  - ACF y PACF muestral.
  - Pruebas de significación individual.
  - Pruebas de significación conjunta (portmanteau).
- Parámetros

Ejemplo



Discusión en clases

Sólo a juzgar por el gráfico de la serie, ¿qué elementos de todos los estudiados hasta el momento reconocen? ¿Tendencia? ¿De qué tipo? ¿Estacionalidad? ¿De qué periodicidad? ¿varianza constante? ¿Valores extremos?

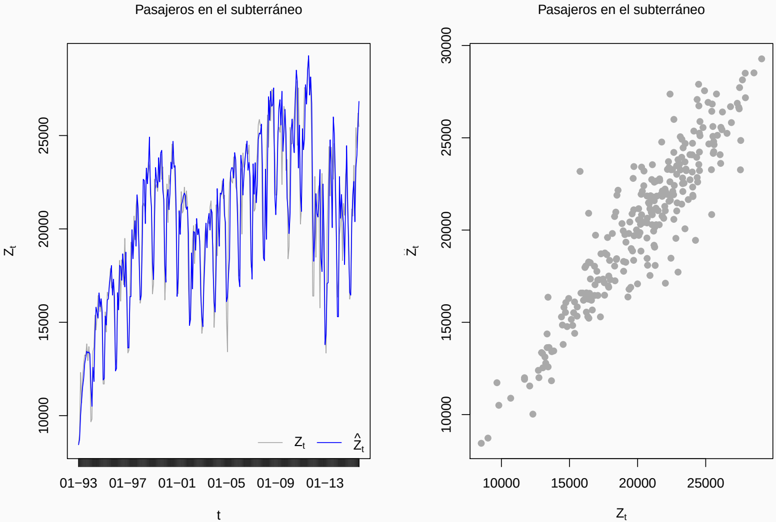
Ajuste

```r
fit <- Arima(
  log(z_ts),
  order    = c(1, 1, 0),
  seasonal = c(2, 0, 0)
)
res <- as.numeric(residuals(fit))

print(fit)
```

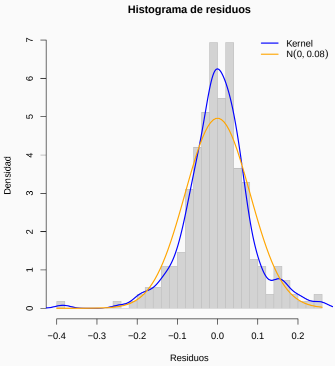
```
## Series: log(z_ts)
## ARIMA(1,1,0)(2,0,0)[12]
##
## Coefficients:
##           ar1    sar1    sar2
##       -0.3569  0.5047  0.3309
## s.e.   0.0575  0.0558  0.0578
##
## sigma^2 estimated as 0.006541:  log likelihood=294.19
## AIC=-580.37   AICc=-580.22   BIC=-565.93
```

Ajuste (continuación)



Distribución Gaussiana

Visualización (1)

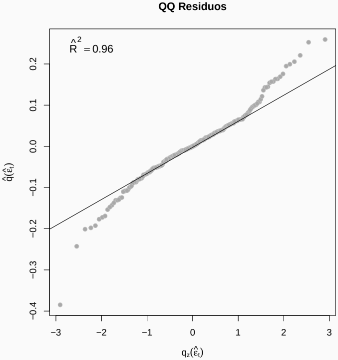



```r
hist(
  res,
  breaks = "FD",
  freq = FALSE
)

lines(
  density(res)
)

curve(
  dnorm(
    x,
    mean = 0,
    sd = sd(res)
    ),
  add = TRUE,
)
```

Visualización (2)




```r
qqnorm(
  res,
  main = "QQ Residuos",
  xlab = expression(
    q[z](hat(epsilon)[t])
    ),
  ylab = expression(
    hat(q)(hat(epsilon)[t])
    ),
  pch = 21,
  bg = "darkgray",
  col = "gray"
  )

qqline(res)
```

Pruebas de hipótesis Jarque-Bera[1]

$$H_0 : a_t \sim \mathcal{N} \wedge \rho_k(a_t) = 0 \;\forall\; k \neq 1$$

$$\frac{n}{6}\hat{S}^2 + \frac{n}{24}(\hat{K} - 3)^2 \overset{H_0}{\sim} \chi^2_2$$

La distribución asintótica converge lentamente. Algunas funciones, como `normtest::jb.norm.test`, calculan el p-value via simulaciones de Monte Carlo.

```r
library(tseries)
jarque.bera.test(res)
```

```
##
##  Jarque Bera Test
##
## data:  res
## X-squared = 71.267, df = 2, p-value = 3.331e-16
```

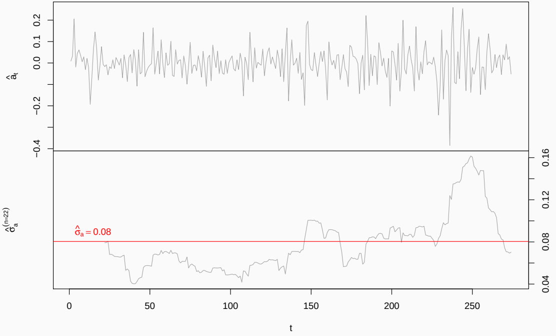
```r
library(normtest)
jb.norm.test(res)
```

```
##
##  Jarque-Bera test for normality
##
## data:  res
## JB = 71.267, p-value < 2.2e-16
```

[1] Otras pruebas relacionadas: Anderson-Darling, Lilliefors, Shapiro-Wilk.

Varianza constante

Visualización



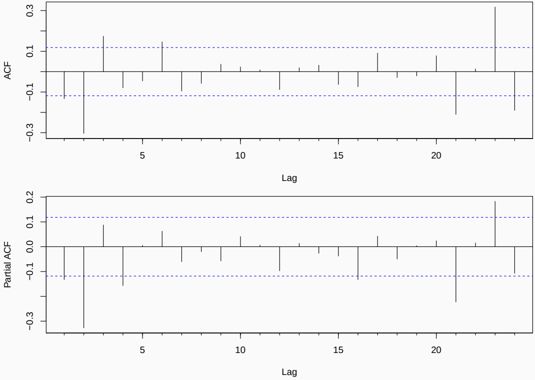

```r
hist(
  res,
  breaks = "FD",
  freq = FALSE
)

lines(
  density(res)
)

curve(
  dnorm(
    x,
    mean = 0,
    sd = sd(res)
    ),
  add = TRUE,
)
```

Independencia condicional

Visualización




```r
par(mfrow = c(2, 1))
Acf(res, main = "")
Pacf(res, main = "")
```

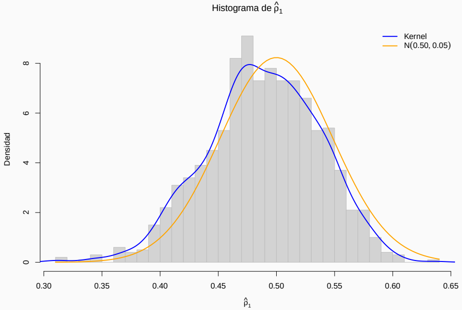
Cálculo del error estándar por simulación

El siguiente es un proceso $AR(1)$,

$$Z_t - 0.5Z_{t-1} = a_t,\ \ t = 0, \pm 1, \ldots,\ \ a_t \sim \mathcal{N}(0, 1).$$

Analíticamente, sabemos que:

$$\gamma_0 = \frac{\sigma_a^2}{1 - \phi_1^2} = \frac{1}{1 - 0.5^2} = \frac{4}{3} \qquad \gamma_k = \phi_1^k \gamma_0 = \frac{4}{3} 0.5^k\ \forall\ k \geq 0 \qquad \rho_k = \phi_1^k = 0.5^k \forall\ k \geq 0 \qquad \mathrm{se}(\hat{\rho}_k) \rightarrow \sqrt{1/T}$$




```r
set.seed(9000)

N    <- 1E3 # Cantidad de simulaciones
T    <- 3E2 # Tamaño de la muestra
phi1 <- 0.5 # Verdadero valor del parámetro
k    <- 1   # Estudiaremos el primer rezago
rho1 <- phi1^k # Valor teórico de la autocorr k=1

rho1sim <- vector("numeric", N)
for (i in 1:N) {
  x <- arima.sim(
    list(ar = phi1),
    n = T
  )
  rho1sim[i] <- cor(head(x, T - k), tail(x, T - k))
}

sprintf(
  "Teórico %0.4f vs. simulado %0.4f",
  sqrt(1 / T), sd(rho1sim)
  )
```

```
## [1] "Teórico 0.0577 vs. simulado 0.0485"
```

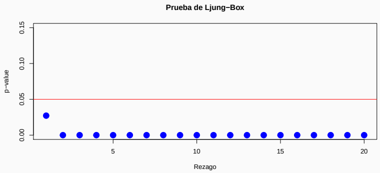
Pruebas de hipótesis Ljung & Box[2]

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_h = 0$$

$$T(T+t)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{(T-k)} \overset{H_0}{\sim} \chi_h^2$$

En lugar de probar individualmente la significancia de cada autocorrelación $\hat{\rho}_k$, la prueba de portmanteau considera conjuntamente la autocorrelación de los primeros h rezagos. Cuidado con la potencia!

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 53.697, df = 12, p-value = 3.095e-07
```




```
Box.test(res, lag = 12, type = "Ljung-Box")

pvals <- sapply(1:20, function(l) {
  Box.test(res, lag = l, type = "Ljung-Box")$p.value
})

plot(
  pvals,
  xlab = "Rezago",
  ylab = "p-value",
  main = "Prueba de Ljung-Box",
  type = "p",
  ylim = c(0, max(pvals, 0.15))
)

abline(h = 0.05)
```

[2] Otras pruebas relacionadas: Durbin-Watson, McLeod-Li, Breusch-Godfrey.

Parámetros

Significatividad de los parámetros

- Alternativa 1: Prueba de hipótesis.
- Alternativa 2: Probar un modelo sin el parámetro. Evaluar diagnóstico y otras medidas de interés (ej. pronósticos).

Selección

Criterios de selección

Luego de muchas transformaciones[3], se arriba a un punto en el proceso del análisis de datos donde un modelo ARMA con media cero resulta suficente. A partir de entonces, nos enfocamos en elegir los órdenes p y q.

La varianza de los pronósticos depende de dos factores: (i) la varianza del error aleatorio, y (ii) la varianza del estimador de los parámetros. Cuando el número de parámetros M crece, el primero se reduce toda vez que el segundo se incrementa. ¿Dónde está el punto justo?

$$AIC = -2 \ln \hat{\mathcal{L}} + 2M \qquad AICc = AIC + \frac{2M^2 + 2M}{T - M - 1}$$

$$BIC = T \ln \hat{\sigma}_a^2 - (T - M) \ln \left(1 - \frac{M}{T}\right) + M \ln T + M \ln \left[\left(\frac{\hat{\sigma}_Z^2}{\hat{\sigma}_a^2} - 1 \right) / M \right]$$

Discusión en clases

¿Hay un punto justo? ¿Qué representan los criterios de información (pista: divergencia de Kullback–Leibler)? ¿Resuelven por completo los problemas de sobreajuste? ¿Por qué?

[3] Estabilizar la varianza, remover la tendencia, remover la estacionalidad

Ejercicio: Ventas en supermercados

Ejercicio en clases

Identificar la serie de tiempo de ejemplo.

El Anexo no incluye la solución :)

Algunos pasos:

- Descargar los datos desde `https://bit.ly/2GXzXoa`.
- De la Sección A 1.11, leer los datos mensuales para la columna *Ventas totales*.[4]
- Ajustar el modelo que propusieron en el ejercicio de la Unidad 6.
- Realizar el diagnóstico de los residuos.

[4] Hay una copia local en `data/INDECSuper.txt` en caso de que el sitio esté fuera de línea.

Ejercicio: Producción de automóviles

Ejercicio en clases

Identificar la serie de tiempo de ejemplo.

El Anexo no incluye la solución :)

Algunos pasos:

- Descargar los datos desde `https://bit.ly/2GXzXoa`.
- De la Sección A 1.22, leer los datos mensuales para la columna *Automóviles*.[5]
- Ajustar el modelo que propusieron en el ejercicio de la Unidad 6.
- Realizar el diagnóstico de los residuos.

[5] Hay una copia local en `data/haciendasAutos.txt` en caso de que el sitio esté fuera de línea.