Series Temporales y Predicción
Práctica 10
Bondad de ajuste y selección del modelo (II)

Glosario

1. Un **buen ajuste a los datos de entrenamiento no siempre** es una indicación de que el modelo **pronosticará bien**.

   Ejemplo: En la referencia [8.10 ARIMA vs ETS | Forecasting: Principles and Practice (otexts.com)](https://otexts.com) encontramos la siguiente salida de R; donde se ha realizado una medición de la bondat de ajuste a un serie realizada con dos modelos distintos: Un mdoelo arima y un modelo ets (suavizado exponencial).

   Para realizar el ajuste a la serie se ha separado la muestra real en dos submuestras: Training set como muestra de entrenamiento de la serie y Test set como muestra para testear si el ajuste predice bien la serie real.

   En la siguiente salida si nos fijamos en los valores que toma el RMSE (raíz cuadrada del MSE donde MSE es el error cuadrático medio calculdoa como la media de los errores al cuadrado) observamos que:

   - Para la muestra Training test el RMSE es menor para el modelo arima que para el modelo ETS
   - Mientras que para la muestra Test set el RMSE es menor en el modelo ETS.

   Así concluimos que el modelo arima ajusta mejor la serie de entrenamiento, mientras que la serie ETS ajusta mejor en las predicciones.

```
# Generate forecasts and compare accuracy over the test set
a1 <- fit.arima %>% forecast(h = 4*(2013-2007)+1) %>%
  accuracy(qcement)
a1[,c("RMSE","MAE","MAPE","MASE")]
#>                RMSE     MAE  MAPE   MASE
#> Training set 0.1001 0.07989 4.372 0.5458
#> Test set     0.1996 0.16882 7.719 1.1534
a2 <- fit.ets %>% forecast(h = 4*(2013-2007)+1) %>%
  accuracy(qcement)
a2[,c("RMSE","MAE","MAPE","MASE")]
#>                RMSE     MAE  MAPE   MASE
#> Training set 0.1022 0.07958 4.372 0.5437
#> Test set     0.1839 0.15395 6.986 1.0518
```

2. La función **coeftest** toma por defecto como distribución del estadístico para los coeficientes del modelo la normal (N(0,1)).

Puesto que el vector de la variable aleatoria de la serie (los residuos) sigue una distribución normal en el modelo arima, el vector de los coeficientes de la variable también lo harán.

Por lo que $b \rightarrow N(\beta, Var(\beta))$. Así, dado el siguiente estadístico sigue $(b - \beta) / S_b \rightarrow t_{n-k}$

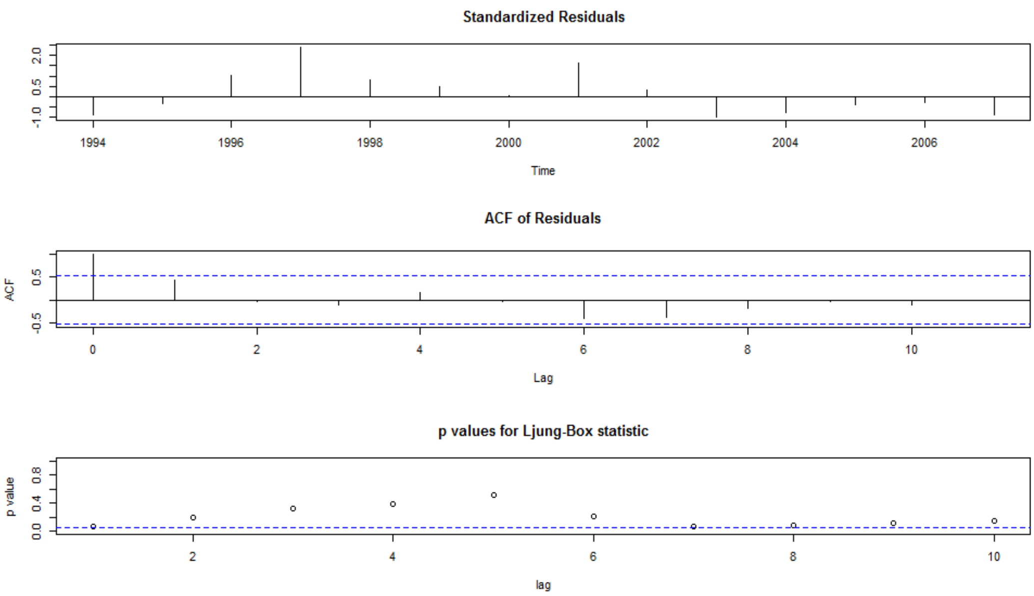
Si n es suficientemente grande $t_{n-k} \rightarrow N(0, 1)$.

Si en la función coeftest se desea que se utilice la t-student para una mejor aproximación cuando la muestra es pequeña, se debe indicar los grados de libertad (df) dentro de la función `coeftest(mod1, df = X)`

3. En el ejercicio 3 de esta práctica realizamos una **regresión donde los residuos siguen un modelo ARIMA**. En el siguiente enlace podéis encontrar más información al respeto: [9.2 Regression with ARIMA errors in R | Forecasting: Principles and Practice (otexts.com)]

Test residuos vistos en esta práctica

4. La función **tsdiag(mod)** nos ofrece la siguiente salida que incluye la visualización de los residuos estandarizados (standarized Residuals), el correlograma de los residuos (ACF of Residuals) y los p-valores del estadístico Ljung-Box.

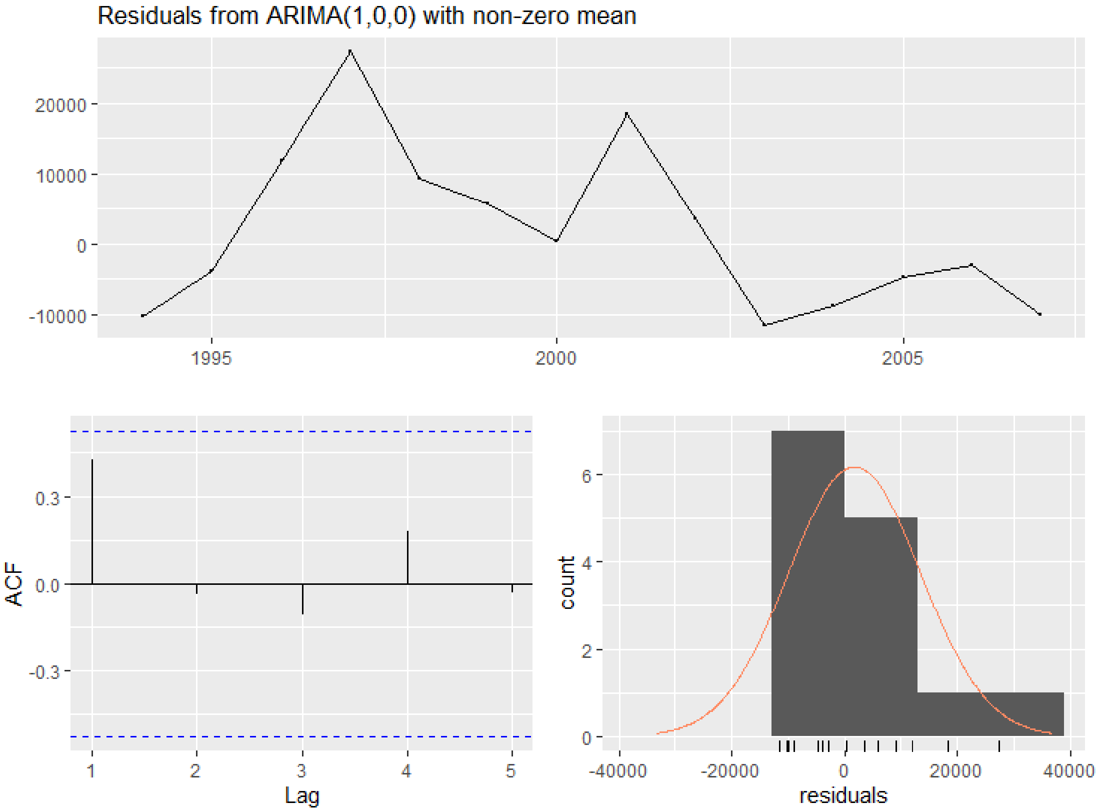


La **prueba de Ljung-Box** se puede definir de la siguiente manera.

H0: Los datos se distribuyen de forma independiente (es decir, las correlaciones en la población de la que se toma la muestra son 0, de modo que cualquier correlación observada en los datos es el resultado de la aleatoriedad del proceso de muestreo).
Ha: Los datos no se distribuyen de forma independiente.

Por lo que el gráfico nos ofrece una prueba de independencia entre los residuos en los diferentes lags. Si todas las bolitas están por encima de la banda, aceptamos la hipótesis de independencia.

5. La funció **checkresiduals(mod)** nos ofrece la siguiente salida que incluye la visualización del gráfico de los residuos, el correlograma de los residuos y el histograma con la curva normal ajustada de los residuos.



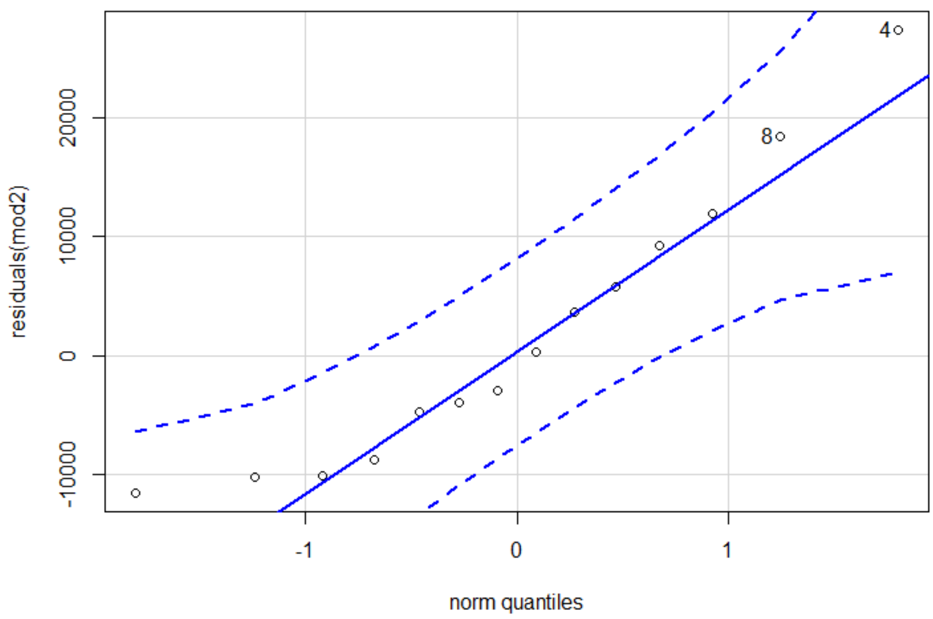
También incluye la **prueba de Ljung-Box** donde se indica el total de lags utilizados.

```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,0) with non-zero mean
Q* = 4.1608, df = 3, p-value = 0.2446

Model df: 2.   Total lags used: 5
```

6. La función Box.test() para realizar el **test Box-Pierce de independencia** de los residuos de la serie.
7. El **gráfico QQPlot** estudia la normalidad de los datos:

   qqPlot(residuals(mod), dist="norm", id=list(method="y", n=2, labels=rownames(Dataset)))

   Cuánto más cerca estan los puntos a la recta diagonal más parecida es la distribución a una normal. Si todos los puntos se encuentran dentro de las bandas nos indica con una confianza del 95% que es una normal.



8. El **test de Kolmogorov-Smirnov** nos mide la normalidad de los datos (H0: Los datos se distribuyen como una normal) viene dado por la función **ks.test**.