Diagnóstico, evaluación y selección de modelos

Curso de Series Cronológicas

Silvia Rodríguez Collazo

Facultad de Ciencias Económicas y de Administración

El diagnóstico del modelo consiste en verificar que las hipótesis realizadas sobre los residuos se cumplen.

Cuando se obtiene una estimación de los parámetros del modelo ARMA, ARIMA o SARIMA en términos más generales, se puede iniciar la siguiente etapa de la modelización, que consiste en verificar el cumplimiento de los supuestos sobre los residuos, en esta etapa se decide si el modelo estimado es estadísticamente adecuado.

El resultado del diagnóstico está vinculado con el proceso de identificación del modelo, dado que si el modelo es inadecuado, será necesario recorrer la etapa de identificación de modo de especificar nuevamente el modelo.

Si como resultado de esta etapa es necesario elegir entre diferentes modelos, los criterios de selección que veremos más adelante nos brindarán elementos para decidir u ordenar estos modelos.

Diagnóstico del modelo

Contrastes vinculados a los parámetros estimados:

- Significación individual de los parámetros
- Observar si hemos sobreparametrizado

Diagnóstico del modelo I

El diagnóstico del modelo requiere comprobar el **cumplimiento de las hipótesis** básicas realizadas respecto a los **residuos del modelo**.

- Media incondicional igual cero
- Varianza incondicional y condicional constantes en el tiempo
- Incorrelación de los residuos para cualquier retardo
- Distribución normal

Estas propiedades deben verificarse no sólo respecto a las distribuciones marginales sino también a las distribuciones condicionadas a cualquier conjunto de información pasada de la serie.

La primer condición es poco restrictiva y aunque se cumpla el modelo puede ser incorrecto, la segunda condición es más fuerte.

La condición de incorrelación de los residuos para cualquier retardo, es central para asegurarse que el modelo es adecuado.

La condición de normalidad es conveniente, porque entre otras cosas garantiza que la incorrelación implique independencia. Cuando los residuos de la serie no se distribuyen normales, es posible que se pueda mejorar el modelo analizando si no hay presentes valores atípicos en la serie(este punto se desarrolla más adelante).

Test de significación individual de las autocorrelaciones estimadas de los residuos

En la práctica no se observan los shocks aleatorios ε_t, sino estimaciones, lo que se tienen son residuos calculados a partir del modelo estimado $\hat{\varepsilon}_t$.

En esta etapa se utilizan los residuos del modelo para contrastar las hipótesis sobre independencia de los shocks aleatorios.

Si los residuos presentan algún tipo de correlación, implica que hay algún patrón que no ha sido considerado dentro del modelo ARIMA estimado, por tanto hay que buscar otro modelo que sí capture esa regularidad.

El primer contraste a realizar es si los residuos estimados están incorrelacionados, de modo de contrastar si cada coeficiente de autocorrelación es significativamente distinto de cero.

Se calcula su función de autocorrelación estimada y si los residuos están incorrelacionados, **los $\hat{\rho}_k$ estimados** (para k no muy pequeño) **serán aproximadamente variables aleatorias con media cero, varianza asintótica 1/T y distribución normal.**

La varianza asintótica es válida para k grande, pero no para los primeros retardos.

Test de significación individual de las autocorrelaciones estimadas de los residuos

Durbin (1970) demuestra que para un proceso AR(1), el desvío asintótico del primer retardo, k=1 de la autocorrelación de primer orden es $\sqrt{(1-\phi^2)/T)}$, que puede ser mucho menor que $1/\sqrt{T}$.

Por tanto el valor $1/\sqrt{T}$ debe considerarse como un límite máximo del desvío de las autocorrelaciones de los residuos.

Se contrasta la hipótesis nula
H_0: $\rho_k = 0$
H_1: $\rho_k \neq 0$
para cada coeficiente de autocorrelación calculando el estadístico t.

Usualmente los programas estadísticos muestran en el gráfico de la FAC estimada y en ella aparecen dos líneas paralelas a una distancia $1/\sqrt{T}$ del origen de las FAC y FACP y por tanto se utiliza como procedimiento habitual para verificar la incorrelación de los residuos.

Pero si el modelo no fuera el correcto,los valores del estadístico Q estarían inflados.

Test de significación conjunta de las autocorrelaciones estimadas de los residuos

Contraste de Ljung-Box

Como forma adicional de verificar que los residuos del modelo están incorrelacionados es a través de la verificación del cumplimiento de esta propiedad mediante una prueba de significación conjunta.

Se contrasta la hipótesis nula
H_0: $\rho_1 = \rho_2 = \rho_3 = \ldots = \rho_h = 0$
H_1: algún $\rho_h \neq 0$
Inicialmente Box y Pierce (1970) proponen el siguiente estadístico para realizar esta prueba:

$$Q_{B-P}(h) = T * \sum_{1}^{h} (\hat{\rho_j})^2 \tag{1}$$

que si el modelo es adecuado, que se distribuye aproximadamente $\chi^2(h-p-q-1)$ grados de libertad, siendo T el número de observaciones disponibles, utilizadas para estimar el modelo.

Contraste de Ljung-Box

Ljung y Box (1978) proponen un contraste global para someter a prueba si las primeras h autocorrelaciones son cero.

Muestran que para tamaños de muestra del tipo usualmente utilizado en la practica, la distribución χ^2 puede no ser una buena aproximación a la distribucion del estadistico Q bajo la hipótesis nula, donde los valores de Q tienden a ser más pequeños de lo que se espera bajo una distribución χ^2. Por lo que proponen un estadístico modificado.

Este estadístico Q(L-B),tiene una $E(Q(L-B))h-p-q$ correspondiente a la distribución $\chi^2(h-p-q)$.

Esta modificación en el estadístico se apoya en que un valor más preciso de la varianza de $\hat{\rho}_k$ es $(T-h)/T(T+2)$ que $1/T$.

Si los residuos siguen un proceso tipo ruido blanco, los coeficientes de correlación estimado son asintóticamente normales, con media cero y varianza $(T-h)/T(T+2)$.

El estadístico de Ljung- Box Q(h)

$$Q_{L-B}(h) = (T(T+2)) * \sum_{1}^{h} (\hat{\rho}_j)^2 / T - j \qquad (2)$$

Concluiremos que el modelo es inadecuado si el valor de Q(h) es mayor que el percentil 0.95 de la distribución χ^2 con (h-n) grados de libertad.

Contraste de media cero

Para contrastar la hipótesis de que las perturbaciones tienen esperanza nula, suponiendo T residuos y $p+q$ parámetros, se calcula su media, su varianza y se plantea una prueba de significación de la media.

Concluiremos que $E(\hat{\varepsilon}) \neq 0$, si

$$\frac{\hat{\varepsilon}}{\hat{s}_{\varepsilon}/\sqrt{T}} \tag{3}$$

es significativamente grande en relación a la distribución $N(0,1)$.

Este contraste debe de aplicarse luego de comprobar que los residuos están incorrelacionados de modo de asegurar que $\hat{s}_{\varepsilon}$ es un estimador razonable del desvío.

Contraste de homocedasticidad

La estabilidad de la varianza incondicional de los residuos se puede analizar estudiando el gráfico de los residuos a lo largo del tiempo. Existen contrastes adicionales para ciertas formas de heterocedasticidad condicional.

En los procesos ARMA tanto la varianza incondicional como la condicional son constantes, pero puede ocurrir que el proceso sea estacionario con varianza incondicional constante y varianza condicional no lo sea.

La varianza condicional representa la incertidumbre en las predicciones, por tanto que la varianza condicional no sea constante implica una incertidumbre en las predicciones variable a lo largo del tiempo.

Este tipo de situaciones se pueden captar observando la FAC y FACP de los residuos al cuadrado (McLeod y Li (1983)). El procedimiento consiste en

estimar las correlaciones de los residuos al cuadrado, crear la FAC y la FACP para esos residuos al cuadrado. Calcular los estadísticos de Ljung - Box correspondientes a los residuos al cuadrado, con este contraste se obtienen elementos para rechazar la hipótesis de varianza condicional constante.

Contraste de normalidad

La normalidad se contrasta con cualquiera de los test de normalidad.

Un contraste posible es el test de Jarque-Bera (1980) (J-B), el estadístico mide la diferencia de los coeficientes de simetría(cs) y curtosis (ck)de la serie con la de una serie con distribución normal, cuyo cs=0 y ck=3.

Se calculan los coeficientes de simetría (cs) y de curtosis (ck) de los residuos y bajo la hipótesis de normalidad, el estadístico:

$$JB = \frac{Tcs^2}{6} + \frac{T(ck-3)^2}{24}$$

sigue una distribución χ^2 con dos grados de libertad.

El coeficiente de simetría de una variable con distribución Normal, es 0 y el coeficiente de curtosis de una variable con distribución Normal, es 3. Si el coeficiente excede el valor de 3, la distribución es más empinada que la Normal (leptocurtica) y si el coeficiente es menor que 3, la distribución es más chata que la normal (platycurtica).

Evaluación de las predicciones

Si el objetivo final de la modelización es la predicción, esta etapa permite comparar entre modelos y conocer el desempeño predictivo del modelo seleccionado.
Existen diferentes formas de evaluar el desempeño predictivo del modelo.

La esperanza condicional es un mejor predictor que la esperanza incondicional?, de no ser así , la información del pasado no estaría contribuyendo, el modelo no contribuye a mejorar la predicción.

Debemos tener en cuenta qué aspectos se quieren evaluar: el desempeño predictivo de la predicción a un paso, a varios pasos, es el mismo modelo el que tiene el mejor desempeño predictivo?

Para evaluar el desempeño predictivo del modelo, se puede seguir los siguientes pasos:

- Se acorta la muestra, por ejemplo dejando fuera una parte de las observaciones (muestra de entrenamiento y muestra para test)
- Se estima el modelo seleccionado y se generan las predicciones correspondientes. La evaluación puede realizarse con las predicciones a un paso y/o a varios pasos.

Evaluación de las predicciones II

- Se calculan los errores de predicción correspondientes usando las observaciones que se dejaron fuera de la muestra (se comparan los valores observados en la muestra test con las predicciones).
- Con esos errores, se calculan los indicadores que se presentan a continuación.
- Si el objeto es comparar el desempeño predictivo de diferentes modelos para los que se verificó previamente el cumplimiento de los supuestos, comparan los indicadores que se presentan más abajo con los errores obtenidos de acuerdo a la utilización de los diferentes modelos.
- Se ordenan los modelos de acuerdo a alguno de estos indicadores:menor error medio, error medio absoluto o menor ECM o raíz del ECM. El primero será el modelo con mejor desempeño de acuerdo a ese indicador.

La elección del criterio que se selecciona para evaluar el desempeño predictivo no es inocua. Recordar las advertencias realizadas en Clements y Hendry(1993) sobre la sensibilidad del indicador ECM respecto a la transformación de la series (niveles, logaritmos, diferencias, diferencias del logaritmo).

Evaluación de las predicciones III

Se pueden construir,a modo de ejemplo, indicadores de desempeño predictivo como: Medidas dependientes de la escala de los datos:

- Error medio $1/F * \sum_{t=1}^{t=F} \hat{\varepsilon}_t$
- Error medio absoluto $1/F * \sum_{t=1}^{t=F} |\hat{\varepsilon}_t|$
- RMSE $1/F * \sum_{t=1}^{t=F} \hat{\varepsilon_t}^2$

Medidas basadas en el procentaje de error:

- Error relativo (relativo al valor observado): $(\hat{e}_t - e_t)/e_t$
- Media absoluta del porcentaje de error (MAPE): media $(|p_t|)$
- Raíz de la media del porcentaje de error al cuadrado (RMSPE) $\sqrt{mean(p_t^2)}$

Siendo F = número de observaciones fuera de la muestra usadas para evaluar la predicción y N= número total de observaciones.

Se recomienda la lectura de Hyndman y Koehler (2006).

Criterios para la selección de modelos

El proceso de identificación puede dar lugar a seleccionar un conjunto de modelos como posibles candidatos, se han elaborado un conjunto de criterios que contribuyan a la selección de modelos.

Contamos entonces con un conjunto de modelos para las observaciones y basados en los datos queremos seleccionar un modelo a partir de un criterio. El criterio de ajuste dentro de la muestra no resulta adecuado pues si comparamos modelos con un número diferente de parámetros, el modelo con más parámetros conducirá a una mayor verosimilitud, por tanto para seleccionar entre modelos se pueden utilizar los criterios de información que se señalan a continuación:

- AIC: Akaike Information Criteria (Akaike 1974)
- AICC: AIC corregido
- BIC: Bayesian Information Criterion (Shwarz 1978)

La idea intuitiva consiste en conciliar la necesidad de minimizar los errores y estimar un modelo parsimonioso (con el menor número posible de parámetros). Estos criterios en general constan de dos componentes, uno que refiere a la minimización de los errores y el segundo, un término de penalización por la incorporación de parámetros adicionales.

Criterio de Akaike (AIC)

$$AIC = T log \hat{\sigma^2_{MV}} + 2 * k \tag{4}$$

Siendo k el número de parámetros estimados para calcular las predicciones a un paso. y $\hat{\sigma}^2_{MV}$ el estimador máximo Verosímil de la varianza de las innovaciones.

Se selecciona el modelo con la verosimilitud esperada máxima.

Pero este criterio tiende a seleccionar modelos con el mayor número de parámetros. Existe un criterio alternativo que corrige este inconveniente y se denomina AIC corregido o AICC.

$$AICc = T log \hat{\sigma^2_{MV}} + T \frac{(1 + k/T)}{1 - (k+2)/T} \tag{5}$$

Si se utiliza el criterio de Akaike para comparar modelos es importante que el número efectivo de observaciones utilizado para estimar los modelos sea el mismo.

Criterios para la selección de modelos

El proceso de identificación puede dar lugar a seleccionar un conjunto de modelos como posibles candidatos, se han elaborado un conjunto de criterios que contribuyan a la selección de modelos.
Contamos entonces con un conjunto de modelos para las observaciones y basados en los datos queremos seleccionar un modelo a partir de un criterio.

- Selección de modelos basado en contraste de hipótesis
- Selección de modelos basado en los errores de predicción
- Selección de modelos basado en criterios de información

Selección de modelos basado en contraste de hipótesis

El modelo se selecciona en base a una secuencia de test. La pregunta es si M_i es preferible al modelo M_{i+1}.

$H_{i0} : M_i$
$H_{ia} : M_{i+1}$

Si H_{i0} no se rechaza, el procedimiento termina allí.

Este procedimiento es útil para la selección del orden de modelo. Para ello, el estadístico de Ratio de verosimilitud puede ser utilizado para la realización del test (Hipótesis anidadas).

En la práctica *forward selection, backward selection y stepwise selection* se utilizan frecuentemente.

Estos procedimientos basados en test de hipótesis están formulados bajo la forma de sucesivos test donde la hipótesis alternativa puede ser incluso una alternativa múltiple pero esto incluye dificultades al procedimiento.

Stepwise regression es un método muy utilizado, pero tiene la desventaja que depende de la trayectoria de búsqueda , *path dependent*.
Se ha mostrado que no tiene una alta tasa de éxito en encontrar el verdadero PGD. Berk (1978) ha demostrado que aplicar tanto *forward* como backward selection no garantiza encontrar el modelo correcto.

Selección de modelos basado en los errores de predicción

Para poder sortear estas dificutades Akaike (1969) propone que el criterio de selección se asocie a los errores de predicción.

Se define el error de predicción final (EPF) como la media al cuadrado de los errores a un paso del modelo ajustado a los datos.

El modelo con el menor EPF es el que se selecciona.Cuando el objetivo final del modelo es hacer predicción, este procedimiento parece ser adecuado.

Sea S_k la suma al cuadrado de los residuos del modelo M_k y sea $\sigma_k^2 = S_k/(n-k)$, usando el estimador insesgado de $\hat{\sigma}_k^2$ se obtiene $\hat{\sigma}_k^2/(n-k)$, que llamamos EFP_k.

Se selecciona el modelo M_{k^*} tal que :

$M_{k^*} = argmin \;\; \text{EFP}_k$

Selección de modelos basado en criterios de información

Sea $z_1, z_2, ...z_n$ n observaciones independientes de un vector de variables aleatorias Z con función de densidad de probabilidad $g(z)$.

Consideremos una familia de funciones de densidad $\{f_\theta(z), \theta \in \Theta\}$.*Siendo*$\theta$ el vector de parámetros y Θ el espacio de parámetros.

Donde la media de la función de verosimilitud está dada por:

$$\frac{1}{n}\sum_{i=1}^{n} logf_\theta(z_i) \tag{6}$$

Cuando n tiende a ∞ esta media tiende a

$$S(g; f_\theta) = \int g(z) logf_\theta(z) dz \tag{7}$$

con probabilidad 1 (se supone la existencia de la integral).
La diferencia

$$K(g; f_\theta) = S(g; g) - S(g; f_\theta) \tag{8}$$

La que se conoce como **distancia de Kullback-Leibler** entre $g(z)$ y $f_\theta(z)$.

Por tanto $S(g; f_\theta)$ se puede utilizar para definir cuál es el modelo que mejor ajusta a través de su maximización.
Maximizar 6 con respecto a θ nos lleva a $\hat{\theta}_{MV}$.

Selección de modelos basado en criterios de información

El criterio de ajuste dentro de la muestra no resulta adecuado pues si comparamos modelos con un número diferente de parámetros, el modelo con más parámetros conducirá a una mayor verosimilitud, por tanto para seleccionar entre modelos se pueden utilizar los criterios de información que se señalan a continuación:

- AIC: Akaike Information Criteria (Akaike 1974)
- AICC: AIC corregido
- BIC: Bayesian Information Criterion (Shwarz 1978)

La idea intuitiva consiste en conciliar la necesidad de minimizar los errores y estimar un modelo parsimonioso (con el menor número posible de parámetros).

Estos criterios en general constan de dos componentes, uno que refiere a la minimización de los errores y el segundo, un término de penalización por la incorporación de parámetros adicionales.

Criterio de Akaike (AIC)

$$AIC(\theta) = -2\hat{\sigma^2}_{MV} + 2k$$

Siendo k el número de parámetros estimados para calcular las predicciones a un paso. Se selecciona el modelo con la verosimilitud esperada máxima.

Este criterio fue designado como una aproximación a un estimador insesgado de la distancia de Kullback-Leibler para el modelo ajustado, el modelo que produce e menor AIC sería la mejor opción.

Pero este criterio tiende sobreparametrizar, a seleccionar modelos con el mayor número de parámetros, a menos que se impongan fuertes restricciones en la dimensión de los candidatos a seleccionar.La imposición de estas restricciones puede ser arbitraria y problemática cuando la muestra es pequeña.

Criterio de Akaike corregido (AICc)

Existe un criterio alternativo que corrige este inconveniente y se denomina AIC corregido o AIC_C y selecciona modeos de menor dimensión.

$$AICc = Tlog\hat{\sigma^2}_{MV} + T\frac{(1+k/T)}{1-(k+2)/T} \tag{9}$$

El AICCc es la suma del AIC más un término de penalización.

Si se utiliza el criterio de Akaike para comparar modelos es importante que el número efectivo de observaciones utilizado para estimar los modelos sea el mismo.

Criterio de Schwarz BIC

Un criterio alternativo, propuesto por Schwarz(1978), desde un enfoque Bayesiano, consiste en maximizar la probabilidad a posteriori del modelo, bajo el supuesto que las probabilidades a priori son las mismas en todos los modelos.

Se puede demostrar que el modelo que maximiza esa cantidad es el que minimiza el criterio.

$$BIC = Tlog\hat{\sigma^2_{MV}} + k * logT \tag{10}$$

donde T es el número efectivo de observaciones, $\hat{\sigma}^2_{MV}$ es el estimador MV de la varianza y k el número de parámetros.

El criterio BIC penaliza más que el AIC por la inclusión de parámetros adicionales, con lo que tiende a elegir modelos más parsimoniosos.

Observaciones finales

- Idealmente tanto el AIC como el BIC deben ser lo más pequeños posibles (ambos pueden ser < 0).
- Como se puede ver a partir de las expresiones de los criterios, para utilizar esto dos criterios sobre modelos alternativos es necesario estimarlos sobre el mismo período (igual muestra), para que sean comparables.
- El criterio AICc y BIC usualmente seleccionan modelos más parsimoniosos que el criterio AIC, ya que el costo de adicionar regresores es mayor.
- Si se utilizan ambos criterios para ordenar modelos, se obtienen resultados diferentes.
- Cuando se usa un único criterio, persiste aun un problema y es cuál es la diferencia aceptable para elegir entre un modelo y otro. Gómez y Maravall (1998) sugieren usar los modelos más balanceados, si los criterios dan valores similares. Ej: entre un ARIMA (2,0,0) y un ARIMA(1,0,1) sugieren elegir el segundo, pues entre otras cosas permite encontrar problemas de factores comunes entre los polinomios de los componentes autorregresivo y de medias móviles.

Referencia bibliográfica

- Box,G.; Jenkins,G.; Rainsel,G; Ljung,G (2016) «*Time Series Analysis.Forecasting and Control.* ».Wiley Series in Probability and Statistics. Fifth Edition.
- Clements,M.; Hendry,D. (1993) «On tehe limitations of Comparing Mean Square Forecast Errors».*Journal of Forecasting*
- Hyndman,R.J.;Koejler,A.B. (2006) « Another look at measures of forecast accuracy ». *International Journal of Forecasting*
- Hubrich,C.;Tsai,CH. (1989) «Regression and time series models selection in small samples». *Biometrika N 76*
- Jarque,C. Bera,A. (1980) « Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals» . *Economic Letters Vol 6. Issue 3.*
- McLeod,A. Li,W. (1983) «Diagnostic Checking ARMA Time Series Models Using Squared Residual Autocorrelations» *Journal of Time Series Analysisi Vol4. N 4*
- Peña, D. (2005) «*Análisis de series temporales*» Alianza Editorial.
- Rao,C.;Wu,Y. (2001) «On Model Selection » IMS Lecture Notes. Monograph Series. Volume 38.