Trabajo Práctico Final - Estadística Actuarial (751-1)

Zajdenberg, Dana Rebeca (892.786)
Barrales Agosti, Julián Guido (889.901)
Cohen Falah, Iván Nahuel (890.693)
Lopatka, Lucas Eitán (889.524)
Borovsky, Ignacio Ariel(891.119)

Introducción

En el presente trabajo se analizarán dos series temporales correspondientes a dos activos financieros distintos, otorgadas por el cuerpo docente. A lo largo del mismo se empleará la metodología de Box-Jenkins, utilizando los conceptos vistos en clase y el programa R-Studio.

Marco teórico

Se define un proceso estocástico como una familia de variables aleatorias que corresponden a momentos sucesivos del tiempo. Será designado por $Y_(t, u)$ donde t es el tiempo y u es la variable aleatoria.

En el siguiente trabajo, se trabajará con una clase especial de los procesos estocásticos: los procesos lineales. Estos procesos se caracterizan porque se pueden representar como una combinación lineal de variables aleatorias.

Son entonces procesos estocasticos lineales:

- El proceso puramente aleatorio Y_t=ϵ_t donde ϵ_t satisface las siguientes propiedades: $E[\epsilon_t] = 0$, $E[\epsilon_t]^2 = \sigma^2$ y $E[\epsilon_t, \epsilon_t] = 0$ En el tratamiento de series temporales, se suele designar a este proceso con la denominacion de "ruido blanco". De ahora en adelante, por ϵ_t se designa únicamente a una variable aleatoria que goza de dichas propiedades

- El proceso autoregresivo de orden p AR(P) se expresa de la siguiente forma,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p}$$

  Al ser p el lag máximo que aparece en el proceso, la denominación de autorregresivo procede de que y_t se obtiene mediante una regresión sobre valores desfasados de la propia variable.

- El proceso de medias móviles de orden q MA(q) dado por,

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ... - \theta_q \epsilon_{t-q}$$

  hace referencia a que la variable y_t se obtiene como un promedio de variables de "ruido blanco", donde las θ resultan ser las variables de ponderación.

- Mediante una combinación de un proceso autorregresivo y un proceso demedias moviles se obtiene un proceso ARMA(p,q), donde p indica el lag maximo de la parte autorregresiva y q señala el correspndiente a la parte de las medias moviles. La expresión de un modelo ARMA es la siguiente:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_q \epsilon_{t-q}$$

En estos procesos estocásticos se supone que cumplen con las condiciones de estacionariedad e invertibilidad. Si se levanta este supuesto el modelo se encuentra en el campo de la no estacionariedad. Para verificar este supuesto se debe cumplir que las raíces del polinomio caracteristico sean mayores a la unidad.

Sea un proceso AR(1) con $\phi_1 = 1$ (es decir, un random walk), se toman diferencias de primer orden para pasar de un proceso y_t a uno w_t

- Todos aquellos modelos que se pueden transformar en estacionarios mediante la toma de diferencias de un determinado orden son los denominados modelos ARIMA(p,d,q)

Metodologia de elaboracion de modelos ARIMA

Hasta ahora, se ha adoptado la perspectiva de un proceso conocido previamente, a partir del cual se han generado una o más realizaciones. El objetivo es, a partir de los valores de una serie temporal, llegar a un conocimiento del mecanismo que verosimilmente haya podido generar la serie.

Para tal fin, se utilizará la el procedimiento que proponen Box y Jenkins. La misma consiste en varias etapas:

1. Fase de identificación
2. Fase de estimación
3. Fase de validación
4. Fase de predicción

Para la **fase de identificación** se identificará el orden de diferenciación. Si se observa la función de autocorrelación, en el caso de que el proceso sea estacionario, decrecerá rapidamente o de manera sinusoidal. Si se cumple el supuesto de que la raíz se aproxima a la unidad, la FAC decrecerá lentamente, lo cual indica que se trata de un proceso no estacionario.

Una vez diferenciada (o no) la serie, se observa la apariencia general de la función de autocorrelación y función de autocorrelación parcial para obtener pistas sobre la elección de los órdenes p y q.

	FACT	FACPT
AR(p)	Decrecimiento rápido de tipo geométrico puro, y geométrico con alteración de signos, sinusoidal o mezcla de varios tipos.	Se anula para lags superiores a "p"
MA(q)	Se anula para lags superiores a "q"	Decrecimiento rápido de tipo exponencial y/o sinusoidal
ARMA (p,q)	Los primeros valores iniciales no tienen patrón fijo y van seguidos de una mezcla de oscilaciones sinusoidales y/o exponenciales amortiguadas.	Los primeros valores iniciales no tienen patrón fijo y van seguidos de una mezcla de oscilaciones sinusoidales y/o exponenciales amortiguadas.

Ahora bien, el comportamiento de la FAC y la FACP da una intuición subjetiva del orden de integración del modelo. Con la finalidad de decidir si el proceso y_t puede ser considerado o no estacionario se utilizará un test de raíz unitaria, particularmente el test de Dickey y Fuller.

La ecuación de Dickey y Fuller es $y_t - y_{t-1} = (\phi_1 - 1)y_{t-1} + \epsilon_t$ donde se testea la hipotesis nula de que $\phi_1 - 1 = 0$ que podria ser contrastada respecto de la alternativa $\phi_1 - 1 < 0$, se utiliza el estdístico Tau. Dickey y Fuller demostraron que la definicion de la zona crítica depende, ademas del tamaño de la muestra, de la forma de la ecuación autorregresiva (por ejemplo, si incluye o no término independente o términos de tendencia determinístca) y definieron tres clases de valores críticos para ser utilizados en los casos que:

- τ: no incluya término independiente ni término lineal
- τ_μ: incluya término independiente, pero no incluya término lineal
- τ_τ: incluya ambos términos

Dado que el estimador de ϕ_1 es positivo se trata de un test de una cola, se considerará que existen razones suficientes para rechazar la hipótesis nula si el valor del estadístico $\tau = \frac{\hat{\phi}_1 - 1}{\hat{\sigma}(\hat{\phi}_1)}$ es menor que el correspondiente valor τ

Luego de identificar la estacionariedad (o no) de la serie se procede a la segunda parte de esta fase. Se debe identificar el orden de la parte autorregresiva (p) y de medias moviles (q). La determinación mas apropiada del orden del modelo se puede lograr con el uso de criterios objetivos de selección de modelos tales como Akaike (AIC), Schwarz (BIC) o Hannan-Quinn (HQ). Este análisis procede con BIC como criterio seleccionado.

La **fase de estimación** consiste entonces en obtener los estimadores de los parámetros del modelo identificado. Existen diversos métodos de estimación para modelos ARMA como el método de mínimos cuadrados y el método de máxima verosimilitud. Se utilizará el primero.

Una vez que se dispone del modelo identificado y estimado, se procede a **fase de validación**.

Análisis de los residuos

Si el modelo especificado es el correcto, y se conocierna los parametros verdaderos y los valores iniciales, se obtiene una serie temporal de "ruido blanco". El analisis deben hacerse entonces a partir de los residuos. Si el comportamiento de los residuos se asemeja al de una serie ruido blanco, existirá una ecuación entre el modelo identificado y los datos.

Con el objetivo de testear si efectivamente los residuos del modelo se asemejan a un ruido blanco se ejecutan dos test:

- *Test de Incorrelación de Ljung-box*: Pone a prueba la incorrelación de los residuos, es decir, a la hipotesis nula de que $\rho_1 = \rho_2 = ... = \rho_M = 0$. A medida que aumenta M disminuye la potencia del contraste.
  Ljung y Box propusieron el siguiente estadístico:

$$Q = T(T+2)\sum_{t=1}^{M}(T-t)^{-1}r_t^2$$

se distribuye como una χ^2 con $M - p - q$ grados de libertad.

- *Test de normalidad de Jarque-Bera*: Pone a prueba la normalidad de los residuos. Bajo la hipótesis nula ϵ_t $Normal()$ contra una alternativa no específica. El estadístico de Jarque-Bera es

$$JB = \frac{n}{4}(S^2 + \frac{(C-3)^2}{4})$$

donde S es el coeficiente de asimetría de los residuos y C el coeficiente de curtosis. El estadístico JB se distribuye como una χ^2 con *2* grados de libertad.

Analisis de los coeficientes estimados

Se contrasta la significatividad de los parámetros del modelo. El estadístico t esta construido sobre la hipotesis nula de que el parámetro es igual a 0. Entonces, para un coeficiente - ϕ_1 por ejemplo - el estadístico t viene dado por la siguiente expresión:

$$t_{N-p-q-\delta} = \frac{\hat{\phi}_1 - (\phi_1/H_0)}{\hat{\sigma}_\phi}$$

Los modelos ARIMA se aplican en general a muestras grandes, por lo que se pueden utilizar los niveles de significación de la normal. Asi, a *groso modo* se rechazaría la hipotesis nula de que $\phi_1 = 0$, para un nivel de significación del 5% cuando $|t^*| >= 2$

Bondad de ajuste

El criterio de Akaike consiste en seleccionar aquel modelo para el que se obtenga un estadístico AIC mas bajo. Este estadístico no presenta el inconveniente que presenta el R^2 y el R^2 corregido, pues penaliza los modelos con mayor número de parámetros y además permite comparar modelos con transformaciones de Box-Cox diferentes.

Reformulación del modelo

Si después de aplicar los contrastes y análisis de los epigráfes anteriores se llega a la conclusión de que el modelo seleccionado no es adecuado, se debe reformular el modelo.

Una vez que se seleccionó y validó el modelo, la **fase de predicción** consiste en en utiilzar este modelo en la predicción de valores futuros de la variable objeto de estudio. Dado un proceso ARMA $\phi(L)Y_t = \theta(L)\epsilon_t$ el valor que se trata predecir se expresa para el valor de Y_{T+l} de la siguiente forma:

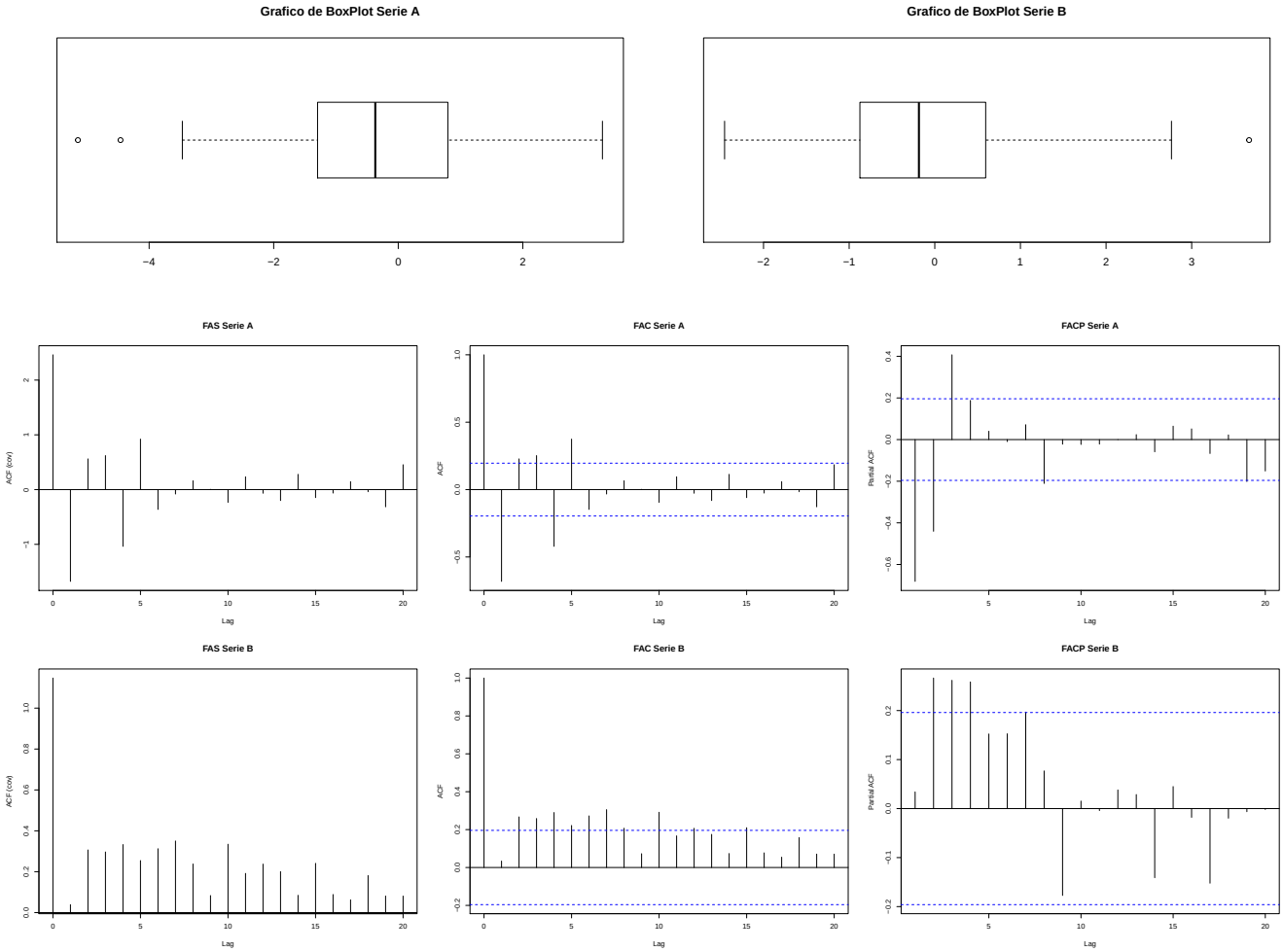
$$Y_{T+l} = \phi_1 Y_{T+l-1} + ... + \phi_p Y_{T+l-p} - \epsilon_{T+l} - \theta_1 \epsilon_{T+l-1} - ... - \theta_q \epsilon_{T+l-q}$$

Para los modelos de medias moviles se cumple que el predictor se anula cuando $l > q$. En general, para el cálculo de predicciones de valores futuros se utiliza el modelo en la forma original ARMA, donde la parte AR esta constituída por coeficientes autorregresivos generalizados.

	Serie A	Serie B
Minimo	-5.1431274	3.2786112
Maximo	-1.2897040	-0.3684681
Percentil 0.25	0.7949836	-0.2597790
Mediana	3.2940126	-0.3189307
Percentil 0.75	2.4864574	1.5768505
Coef Asimetria	-5.1431274	3.2786112
Coef Curtosis	-1.2897040	-0.3684681
Promedio	0.7949836	-0.2597790
Varianza	3.2940126	-0.3189307
Desvio	2.4864574	1.5768505

Fase de identificación

Se grafican la series y sus funciones de autocovarianza y autocorrelacion para una primera impresión



Para la serie A, el gráfico de la FAC muestra un decrecimiento sinusoidal. En cuanto a la serie B, a través de la FACP se verifica que decrece lentamente de manera sinusoidal. Los gráficos proponen la posibilidad de que la serie B sea un proceso de medias moviles no estacionario. Para confirmar estas intuiciones son realizados los correspondientes tests de Dickey-Füller con el objetivo de determinar la existencia o no de raíces unitarias.

Los valores de Tau empíricos para cada modelo son:

	Serie A	Serie B
None	-1.493122	-2.915783
Con Drift	-3.133971	-1.493122
Con Trend	-2.915783	-3.133971

Y los respectivos valores criticos de Tau son

	1pct	5pct	10pct
None	-2.60	-1.95	-1.61
Con Drift	-3.51	-2.89	-2.58
Con Trend	-4.04	-3.45	-3.15

Para la serie A se puede afirmar con un nivel de significatividad aproximado a 5% que la serie es no estacionaria para todos los modelos. En cuanto a la serie B, existe evidencia suficiente para no rechazar la hipótesis nula de estacionariedad para modelos *none* y *drift*. Sin embargo, se rechaza para un modelo con tendencia. (Supone un α del 5%)

Fase de estimacion

A continuacion, se evaluaron todas las posibles combinaciones de modelos ARIMA (p,d,q) con p, d y q entre cero cuatro y tres con el fin de seleccionar aquel que mejor ajusta en base a los criterios de informacion vistos en clase.

```
## [1] "Para la serie A, el menor AIC fue de 267.636093520523 para el modelo ARIMA(4, 0, 4)"
## [1] "Para la serie A, el menor BIC fue de 282.056862231562 para el modelo ARIMA(2, 1, 1)"
## [1] "Para la serie B, el menor AIC fue de 274.510249397882 para el modelo ARIMA(0, 1, 2)"
## [1] "Para la serie B, el menor BIC fue de 282.295608948286 para el modelo ARIMA(0, 1, 2)"
```

Para la Serie A, el modelo que mejor se ajusta segun el criterio BIC es un ARIMA (2,1,1).

$$\hat{y_t} = -1.2741 y_{t-1} - 0.7230 y_{t-2} + 0.6280 \epsilon_{t-1}$$

Para la Serie B, el que mejor se ajusta segun BIC es el modelo ARIMA (0,1,2)

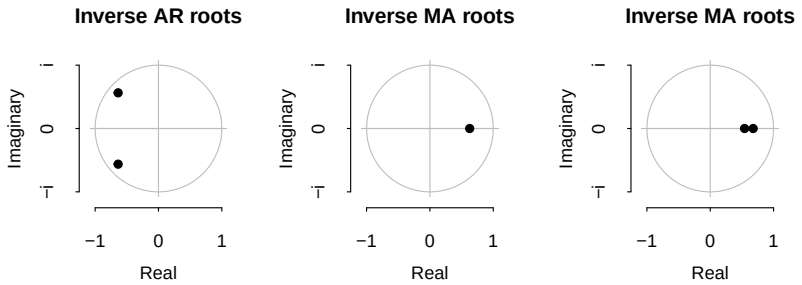
$$\hat{y_t} = 1.2202 \epsilon_{t-1} - 0.3677 \epsilon_{t-2}$$

Fase de validación

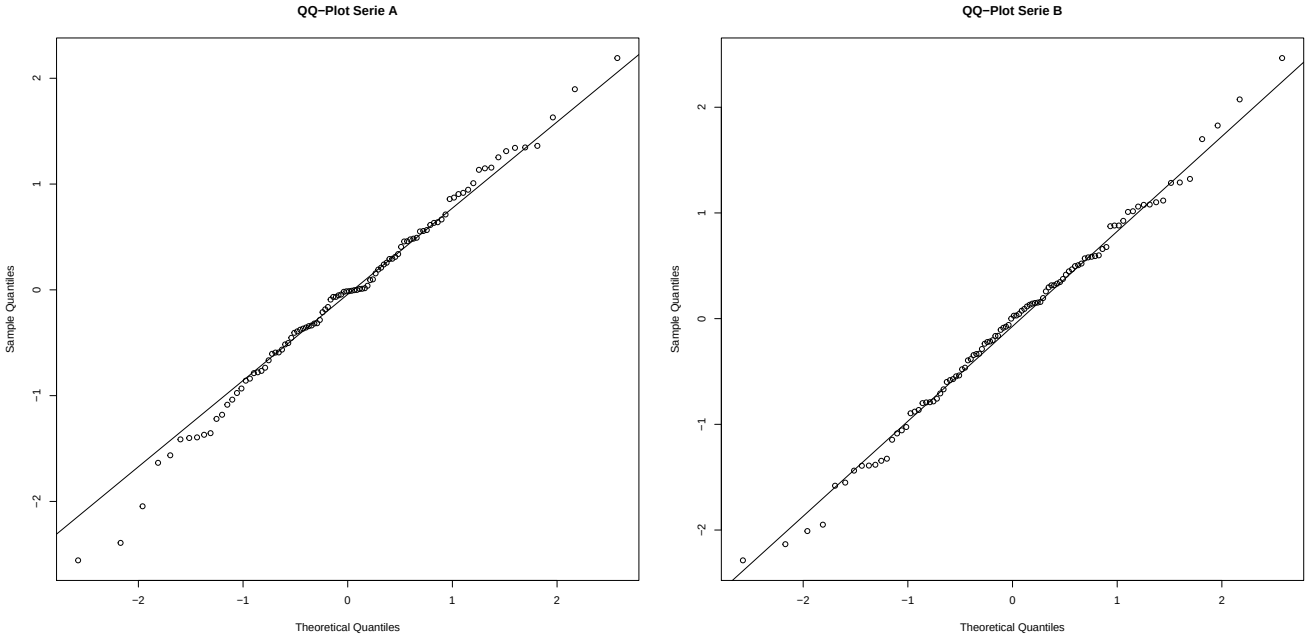
Se demuestra primero que los coeficientes del modelo estimado son significativos. Para esto es utilizada una función que devuelve un mensaje. Se ingresan los modelos estimado como input y el codigo devuelve

```
## [1] "El coeficiente ar1 es un parametro significativo"
## [1] "El coeficiente ar2 es un parametro significativo"
## [1] "El coeficiente ma1 es un parametro significativo"
## [1] "El coeficiente ma1 es un parametro significativo"
## [1] "El coeficiente ma2 es un parametro significativo"
```

El siguiente paso es que los modelos cumplan con las condiciones de estacionariedad e invertibilidad.



Las raices caen dentro del circulo unidad. Ahora se analiza la normalidad de los residuos. Para ello se realiza un gráfico de *qqnorm* que compara los cuantiles teóricos de la distribución normal con los de los residuos de la serie.



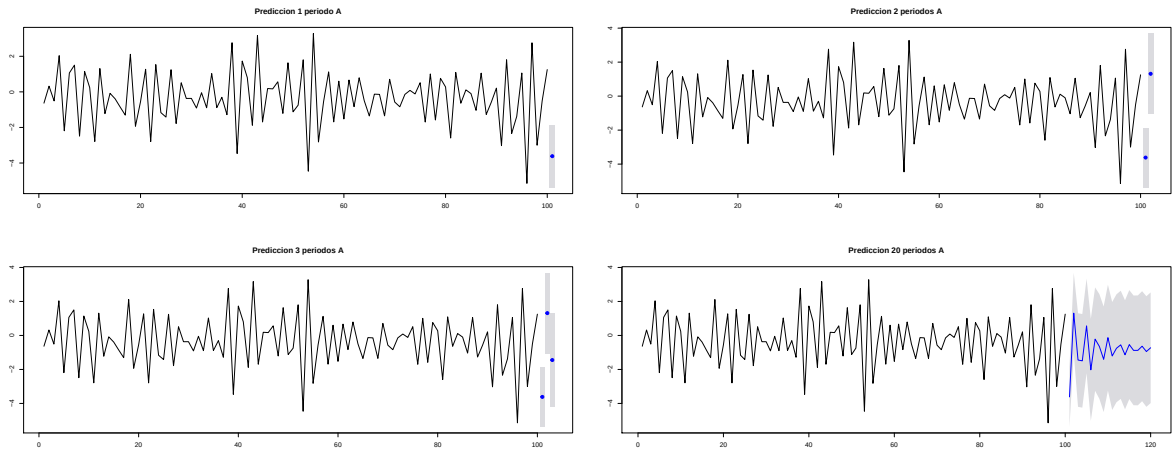
Se observa que los cuantiles teóricos coinciden con los cuantiles de los residuos. Para complementar el **analisis de normalidad** es realizado un test de Jarque Bera

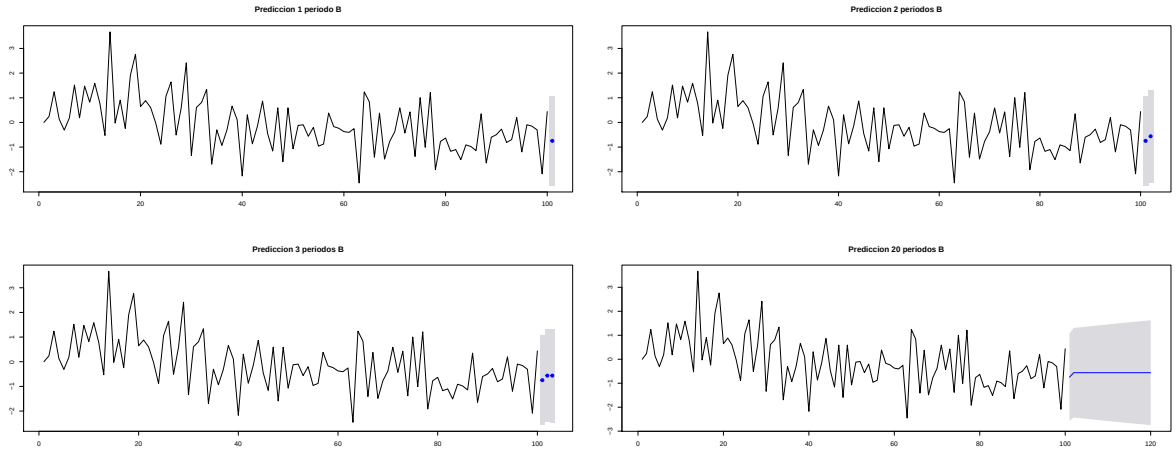
Como los *p-value* de ambas series son considerablemente mayores a un α de 0.1 no se rechaza la hipotesis nula de normalidad de los reisudos. Por ultimo, la **incorrelacion de los residuos** es analizada. Se ejecuta el test de Ljung-Box y se obtiene,

Los test de incorrelación para rezagos de 4, 6, y 8 arrojan respectivamente valores p mayores a un α de 0.1. Por lo tanto, hay suficiente evidencia de que los residuos de ambos modelos estan incorrelacionados para valores defasados.

Fase de Predicción

Para la serie B no es correcto predecir a partir de un modelo de medias móviles. Sin embargo, se ejecuta el código para observar los resultados. La predicción finalmente para un horizonte de uno, dos, tres y veinte periodos y con un nivel de confianza al 95% es.





Los valores de los gráficos son representados en los siguientes cuadros. Se calcularon los intervalos para el 94%, 95% y 99% de confianza.

Table 2: Intervalos de confianza para el Modelo A

	Limite inferior 0.99	Limite inferior 0.95	Limite inferior 0.94	Prediccion	Limite sueprior 0.94	Limite superior 0.95	Limite superior 0.99
1	-5.926600	-5.374150	-5.3031319	-3.616003	-1.928873	-1.857855	-1.305405
2	-1.798766	-1.054716	-0.9590676	1.313188	3.585444	3.681092	4.425142
3	-5.064972	-4.199999	-4.0888049	-1.447258	1.194290	1.305483	2.170457
20	-5.002368	-3.979625	-3.8481501	-0.724792	2.398566	2.530041	3.552784

Table 3: Intervalos de confianza para el Modelo B

	Limite inferior 0.99	Limite inferior 0.95	Limite inferior 0.94	Prediccion	Limite sueprior 0.94	Limite superior 0.95	Limite superior 0.99
1	-3.144299	-2.571402	-2.497756	-0.7481834	1.001389	1.075036	1.647932
2	-3.015761	-2.429135	-2.353723	-0.5622211	1.229281	1.304693	1.891319
3	-3.041058	-2.448383	-2.372194	-0.5622211	1.247752	1.323941	1.916616
20	-3.437248	-2.749847	-2.661480	-0.5622211	1.537038	1.625404	2.312806

Efectuado el analisis, se procede a exportar los datos en formato CSV

Conclusiones

Se realizaron dos análisis, uno por cada serie.

Ambas series resultaron ser no estacionarias y validaron todos los supuestos vistos en la materia, para luego concluir con una predicción para 20 periodos. Se modelizó la primer serie para un ARIMA(2,1,1) y la segunda para un ARIMA(0,1,2). Se observa que el activo financiero correspondiente a la serie A tiene un comportamiento que a lo largo del tiempo tiende al cero. Por lo tanto la inversión en este activo financiero no sería la más adecuada. De todos modos, se debería tener en cuenta que la predicción para un horizonte mayor a 2 en un modelo ARIMA(2,1,1) conlleva un error significativo. Se observa que el activo financiero correspondiente a la serie B tiene un comportamiento que una vez superado el orden de la parte MA del modelo, es constante y aproximadamente cero. Por lo tanto la inversión en este activo financiero no sería adecuada. De todos modos, se debería tener en cuenta que al realizar la predicción con un modelo MA se obtienen resultados muy poco certeros, con una gran carga de error. Es por ello que la decisión de invertir o no en dicho activo no puede basarse en esta predicción, se debrían buscar otros modelos que tengan unna mayor capacidad predictiva.

Futuras investigaciones y limitaciones

El análisis efectuado se limitó a los modelos de la familia ARIMA. No se tomó en cuenta la posibilidad de realizar una diferencación fraccionaria, ni modelizar la serie en otras familias de modelos para series de tiempo.

Por otra parte, se utilizaron los tests vistos y estudiados en la materia *751-Estadistica Actuarial*. Como se utilizó Dickey-Fuller para el análisis de raíces unitaras se pudieron haber utilizado otros tests como el de KPSS o Phillips-Perron.

Bibliografía

- *Analisis de series temporales: modelos ARIMA* - Ezequiel Uriel Jimenez; Valencia, España, 1985.
- *"Time Series Analysis. Forecasting and Control. Fourth edition"* - George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel.
- *Acerca de la probabilidad* - Alberto H Landro, 2010.

Anexos

Todo el código R empleado para realizar este analisis está disponible en el siguiente enlace