Estadística Actuarial

Modelos de series de tiempo

1. Introducción y contenidos

En el presente estudio se buscó desde un primer enfoque adentrarse en la modelización de series temporales. Para ello será necesario introducir algunos conceptos fundamentales para la correcta comprensión del estudio. Se explicará conceptualmente el significado de *proceso estocástico* lo cual nos dará pie para comprender lo que es una serie temporal e introducir conceptos como estacionariedad. Veremos que para comprender el correcto funcionamiento de un proceso estocástico es necesario contar con conocimientos estadísticos previos. En nuestro caso se mostrará una aplicación práctica en la que se utilizarán las funciones de *autocorrelación y autocorrelación parcial* buscando comprender más en profundidad la serie. Veremos un primer enfoque a la modelización de series de tiempo, más específicamente, la cantidad de casos positivos de COVID-19 con *modelos autoregresivos* (AR), *modelos de medias móviles (MA) y* la combinación de ambos ARMA.

2. Proceso estocástico

Un proceso estocástico se puede representar como una sucesión de variables aleatorias $\{X_t : t \in T\}$ definidas en el mismo espacio muestral. La variable T es el tiempo la cual dependiendo del proceso puede ser una variable continua $(\mathbb{R}; \mathbb{R}^n; \mathbb{R}_{zo})$ o discreta $(\mathbb{N}; \mathbb{N}_o; \mathbb{Z})$. Desde un punto de vista más intuitivo podríamos caracterizar a los procesos estocásticos como una secuencia de variables que evoluciona en el tiempo. Teniendo en cuenta un proceso estocástico finito $X(\omega, t)$ es decir que $\{t = 0,1; 2; 3; \ldots n\}$ podemos representar su función de distribución conjunta finita como:

$$F_{X_{t_1};X_{t_2};\ldots X_{t_n}}(a_1; \ldots; a_n) = P(\omega : X_{t_1} \leq a_1; \ldots; X_{t_n} \leq a_n) \qquad (1)$$

2.1. Estacionariedad

Se dice que un proceso es estacionario de primer orden cuando su función unidimensional es decir, no conjunta, no varia en el tiempo. Diremos en cambio que un proceso de orden n es estacionario si su distribución conjunta finita de n variables no varia con el tiempo. Para ello debe cumplir con la siguiente relación:

$$F_{X_{t_1};X_{t_2};\ldots X_{t_n}}(a_1; \ldots; a_n) = F_{X_{t_{1+k}};X_{t_{2+k}};\ldots X_{t_{n+k}}}(a_1; \ldots; a_n) \qquad (2)$$

Si cumple con la relación de la ecuación (2) podemos afirmar que el proceso es estacionario en sentido fuerte. Sin embargo existe también la posibilidad de que el proceso sea estacionario en sentido débil. Eso sucede cuando no es posible afirmar que la distribución conjunta finita no varia

en el tiempo, pero sí podemos afirmar que su *esperanza, varianza y covarianza* no varian dependen del tiempo.

$$E(X_t) = E(X_{t+k})$$
$$V(X_t) = V(X_{t+k})$$
$$Cov(X_{t_1}; X_{t_2}) = Cov(X_{t_1+k}; X_{t_2+k})$$

2.3. Ruido blanco

Un proceso Z_t se lo llama ruido blanco si esta compuesto por una serie de variables aleatorias independientes e idénticamente distribuidas. El mismo debe tener una media constante a lo largo del tiempo al igual que su varianza, por otro lado por ser variables aleatorias independientes, la covarianza entre dos variables cualesquiera de un proceso de ruido blanco debe ser igual a 0.

$$E(\varepsilon_t) = \mu_\varepsilon \quad \forall t \in T$$
$$V(\varepsilon_t) = \sigma_\varepsilon^2 \quad \forall t \in T$$
$$Cov(\varepsilon_t; \varepsilon_{t+k}) = 0 \quad \forall (t,k) \in T$$

2.4 Procesos autoregresivos

Los procesos autoregresivos son aquellos que se basan en que la variable dependiente o regresada, puede ser explicada por la misma variable pero en momentos anterior $(t-1)$. Se los denota como $AR(p)$ donde p indica la cantidad de valores pasados utilizados para explicar la variable dependiente en el momento t. Por lo que un modelo autoregresivo de orden p sería:

$$AR(p) \qquad X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t$$

Donde los ϕ_i son los parámetros del proceso/modelo y el ε representa una variable de ruido blanco pertinente al momento de la variable explicada.

2.5 Medias móviles

Otra alternativa de formar procesos univariados es a travez de los procesos de medias móviles. Los mismos son útiles para modelizar procesos en los cuales los eventos que suceden dependen más fuertemente con su pasado cercano que con el pasado lejano. Se denotan con MA(q) por sus siglas en inglés. Podemos expresar un proceso de medias móviles de orden q de la siguiente manera:

$$MA(q) \qquad X_t = \mu + \theta\varepsilon_t + \theta\varepsilon_{t-1} + \ldots + \theta\varepsilon_{t-q}$$

Vemos que el orden del proceso indica la cantidad de periodos pasados que son considerados. En cuanto a las variables μ es la media del proceso, θ es parámetro y ε_i es ruido blanco.

3. Función de autocorrelación

La función de autocorrelación nos indica qué tan relacionada esta el proceso estocástico con respecto a una versión desplazada del mismo. Esta función nos permite obtener detalles los cuales pueden ser de utilidad para encontrar patrones dentro del proceso. La misma se calcula de la siguiente manera:

$$\rho_k = \frac{Cov(X_t; X_{t+k})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t+k})}}$$

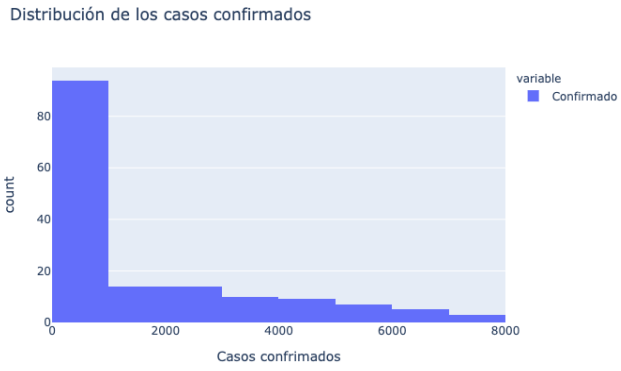
Notar que en el caso de que el proceso sea estacionario ambas varianzas serán iguales por lo que la función de autocorrelación sería $\rho_k = \frac{\gamma_0}{\gamma_1}$ dado que $Var(X_t) = Var(X_{t+k}) = \gamma_1$.

3.1 Función de autocorrelación parcial

La función de autocorrelación parcial hace referencia a la correlación directa que hay entre dos variables separadas en el tiempo. Es decir, al calcular la autocorrelación entre dos variables $X_t; X_{t+k}$ del proceso, dentro se incluyen las relaciones lineales de las variables $X_{t+1}; X_{t+2}; \ldots; X_{t+k-1}$. En cambio al calcular la función de autocorrelación parcial se estarían dando por dadas dichas relaciones teniendo en cuenta únicamente la relación directa entre $X_t; X_{t+k}$.

4. Anexo: modelización COVID-19

Dados los contenidos teóricos que considere fundamentales y pertinentes a modelización de una serie temporal, presentaré un estudio realizado sobre los casos positivos de COVID-19 en Argentina. Ya que es una tarea difícil, el resultado que presentado no es más que un primer acercamiento a la modelización de dicha serie de tiempo, como verán al final con los resultados queda mucho por mejorar. Sin embargo considero importante y necesario exponer el análisis realizado durante el estudio. La base de datos utilizada fue utilizada es la que provee el Ministerio de Salud de Argentina. El software utilizado para el análisis y modelización de los datos es el lenguaje Python.



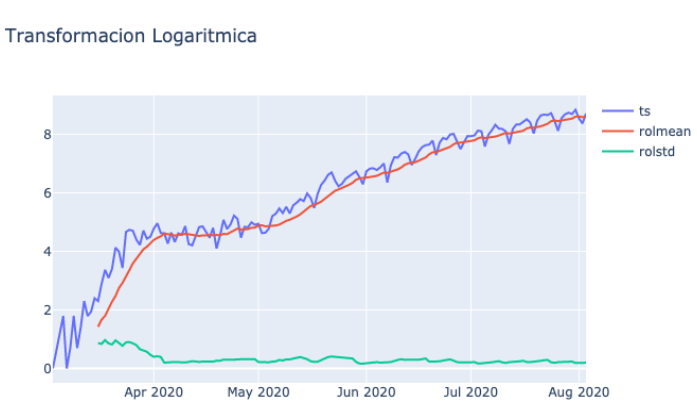
En primera instancia se visualizó la serie temporal y la distribución de la misma, una vez que sabemos la distribución vamos a poder sacar conclusiones acerca de la serie, por ejemplo analizar la estacionariedad. Para ello se utilizaron dos métodos en simultáneo, por un lado se buscaba encontrar atributos que evidenciaban la no estacionariedad, en caso de no encontrarlos se procedía a hacer el Dickey-Fuller test. El mismo plantea la hipótesis nula de que existe una raíz que sea $r < |1|$, de esta manera si la hipótesis nula se

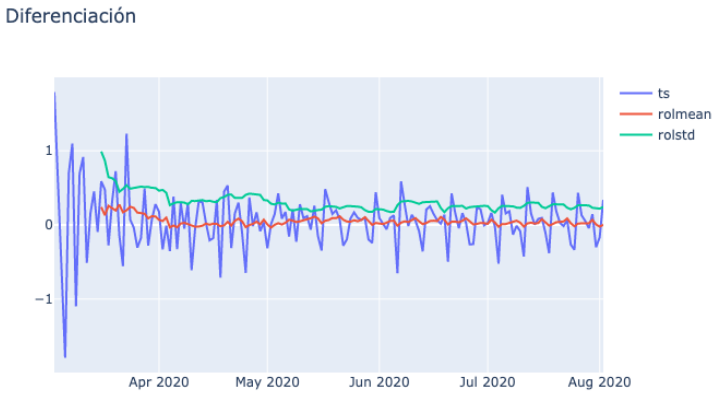
rechaza y no existe una raíz menor a uno en valor absoluto podríamos concluir que es estacionaria.

Se comprobó que la serie sigue una distribución exponencial, lo cual nos podría generar problemas. Ahora sabemos que la serie de tiempo tal y como esta no representa un proceso estacionario. Para ello va a ser necesario hacerle algunas transformaciones.

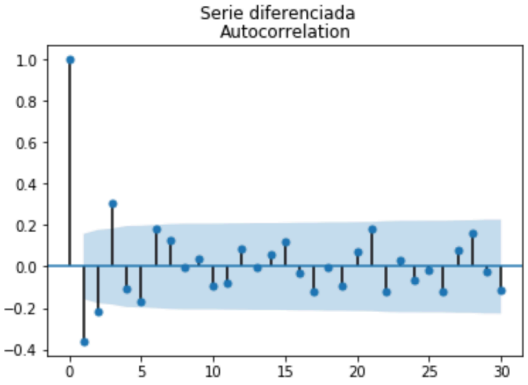
En primera instancia se realizó una transformación logarítmica. Una vez hecha la transformación se realizó el Dickey-Fuller test. El p-valor dio 0.54, por lo que aún luego de la transformación logarítmica no fue suficiente para hacer de la serie un proceso estacionario.

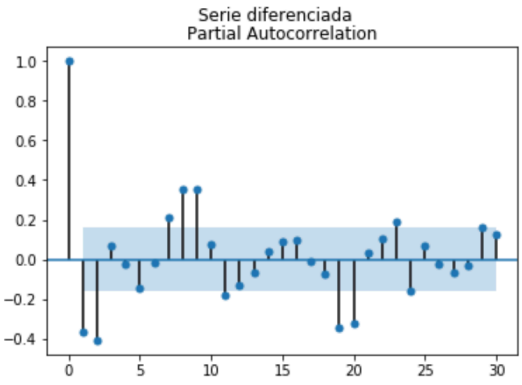
Dado que la transformación logarítmica no alcanzó, se aplicó otro tipo de transformación, diferenciación. Luego de diferenciar se volvió a probar el Dickey-Fuller test. Esta vez si se logró, el p-valor arrojado fue de 0.01 por lo que con un $\alpha = 0.05$ podemos afirmar que nuestra serie es estacionaria.





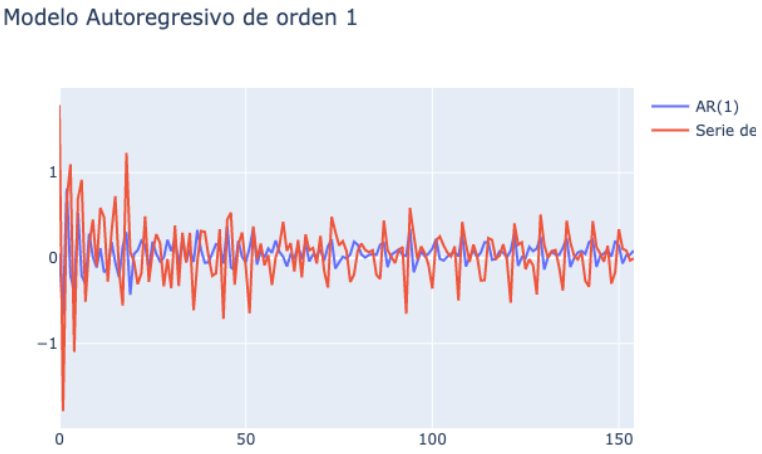
Recordando las funciones de autocorrelación y autocorrelación parcial anteriormente vistas, procedimos a calcularlas y visualizarlas en busca de algún patrón dentro de la serie de tiempo. De esta manera sería más fácil la elección de los parámetros de los modelos. A continuación se muestran los gráficos de autocorrelación y autocorrelación parcial para la serie de tiempo luego de la transformación logarítmica y de la diferenciación.

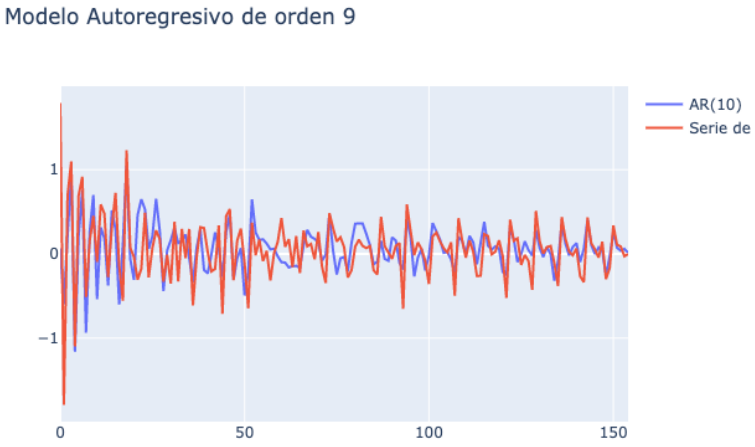




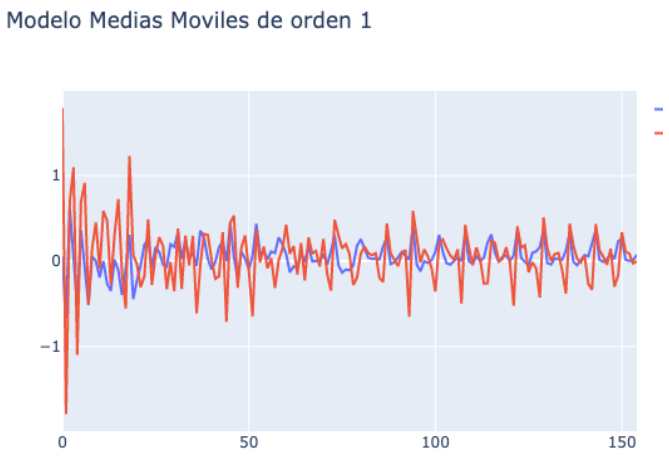
En primera instancia se ve que el carácter de las funciones siguen cierto patron sinusoidal. A su vez una buena práctica podría ser relacionar el orden del modelo autoregresivo con el último lag con autocorrelación significativa, es decir que supera el desvío estándar. Sin embargo eso no fue posible ya que no existe una convergencia de las autocorrelaciones parciales. Por otro lado lo mismo podría asociarse entre la autocorrelación y el modelo de medias móviles. En este caso vemos que si podría decirse que las autocorrelaciones a partir del 3er lag convergen dentro de los desvíos estándar. Sin embargo al probar para distintos valores del modelo de medias móviles se comprobó que le de orden 3, no es el que mejor performa. Para ello se utilizó el calculo de RMSE entre la serie modelizada y la real.

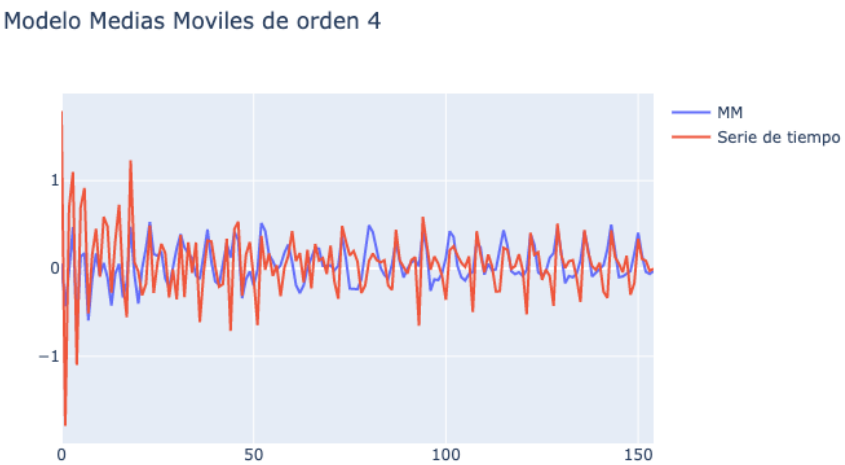
Vemos a continuación un primer enfoque al modelo autoregresivo en el cual se itero con distintos ordenes y se vio que a medida que aumentaba el orden el RMSE converge. A fines prácticos se muestra una visualización de los AR(1) y AR(9):



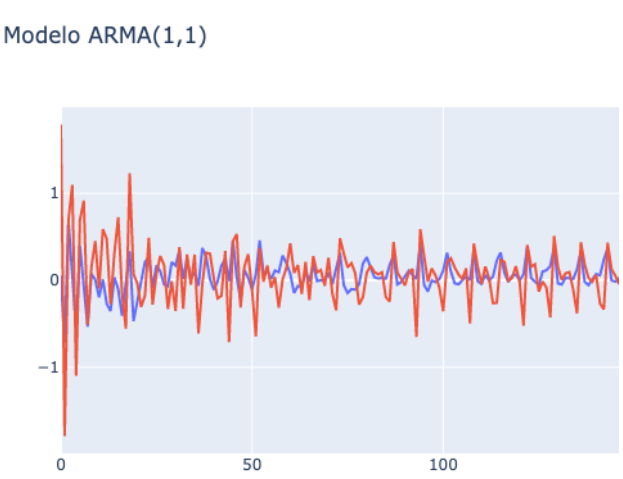


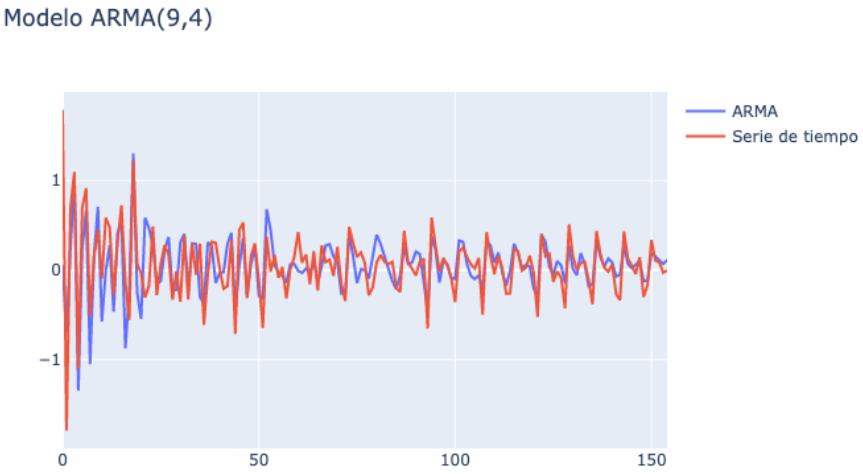
En la siguiente iteraciones se realizó en cambio el modelo de medias móviles, en este caso de orden 1 y 4, es decir MA(1) y MA(4):





Por último se realizó una combinación de ambos modelos, es decir el modelo ARMA. Primero en su expresión más básica es decir ARMA(1,1) y luego con ARMA(9,4):





Entre todos los modelos que se probaron aquel que mejor resultó fue el de ARMA(9,4) con un RMSE igual a 0.08314110696496035

Finalmente se eligió la serie del modelo que mejor resultó, para ello se evaluó el resultado de los RMSE. Luego se le aplico la inversa de sus transformaciones para poder comprarla con la serie real. El resultado fue el siguiente: