MODELIZACIÓN DEL COVID-19 EN SANTANDER MEDIANTE SERIES TEMPORALES

CRISTIAN JULIÁN DÍAZ GARCÉS

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2023

MODELIZACIÓN DEL COVID-19 EN SANTANDER MEDIANTE SERIES TEMPORALES

CRISTIAN JULIÁN DÍAZ GARCÉS

Trabajo de grado para optar al título de
Matemático

Director
Andrés Sebastián Ríos Gutiérrez
Candidato a doctor

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2023

DEDICATORIA

Este trabajo viene dedicado para todas las personas que me ayudaron a lo largo de mi carrera y a la elaboración de este trabajo, también al esfuerzo de nunca rendirse y a las personas que indirectamente ayudaron a mi bienestar en mi estadía en la universidad.

AGRADECIMIENTOS

Agradezco a mi familia por el apoyo económico que tuvieron para conmigo durante el desarrollo de mi carrera. También agradezco a mis amigos que me ayudaron en los momentos donde pensé en darme de baja de la carrera y a los profesores de la escuela que son pocos los que me ayudaron a considerar terminar la carrera.

Un reconocimiento y agradecimiento importante a mi director de trabajo de grado, por que sin conocerme me presto su tiempo y sus conocimientos en todo momento y me apoyo en todo lo que necesite, le estoy profundamente agradecido.

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

RESUMEN

TÍTULO: MODELIZACIÓN DEL COVID-19 EN SANTANDER MEDIANTE SERIES TEMPORALES [*]

AUTOR: CRISTIAN JULIÁN DÍAZ GARCÉS [**]

PALABRAS CLAVE: SERIES DE TIEMPO, MODELIZACIÓN, COVID-19, PREDICCIÓN, ARIMA MULTIPLICATIVO, GARCH, TEORÍA DE PROBABILIDAD, ESTADÍSTICA.

DESCRIPCIÓN:

En los últimos años la pandemia del COVID-19 cambio mucho la vida como la conocíamos, conocer como esta pandemia se propagaba y lograr conocer los posibles contagiados es muy importante para tomar decisiones de salud publica para evitar el colapso del sistema de salud. Este trabajo consiste en presentar un modelo ARIMA para la predicción del número de casos y de ser necesario un modelo GARCH si los errores del modelo ARIMA no se comportan de buena manera (Homocedasticidad).

En el primer capítulo, recordaremos algunos conceptos importantes de teoría de series de tiempo ARIMA, y todo lo relacionado a su modelización. En el capítulo siguiente presentaremos algunas definiciones de los modelos GARCH, como se realiza su estimación. En el ultimo capitulo veremos una simulación para confirmar nuestra metodología planteada y el análisis de los datos de COVID-19.

---

[*] Trabajo de grado

[**] Facultad de Ciencias. Escuela de Matemáticas. Director:Andrés Sebastián Ríos Gutiérrez, Candidato a doctor.

ABSTRACT

TITLE:MODELING OF COVID-19 IN SANTANDER THROUGH TIME SERIES [*]

AUTHOR: CRISTIAN JULIÁN DÍAZ GARCÉS [**]

KEYWORDS: TIME SERIES, MODELING, COVID-19, PREDICTION, MULTIPLICATIVE ARIMA, GARCH, PROBABILITY THEORY, STATISTICS.

DESCRIPTION: In recent years, the COVID-19 pandemic has changed life as we knew it a lot. Knowing how this pandemic spread and getting to know the possible infected people is very important to make public health decisions to avoid the collapse of the health system. This work consists of presenting an ARIMA model for the prediction of the number of cases and, if necessary, a GARCH model if the errors of the ARIMA model do not behave well (Homoscedasticity).

In the first chapter, we will recall some important concepts of ARIMA time series theory, and everything related to its modeling. In the next chapter we will present some definitions of the GARCH models, how their estimation is carried out. In the last chapter we will see a simulation to confirm our proposed methodology and the analysis of the COVID-19 data.

---

[*] Bachelor Thesis

[**] Facultad de Ciencias. Escuela de Matemáticas. Director: Andrés Sebastián Ríos Gutiérrez, Candidato a doctor.

Introducción

La predicción de comportamientos, ya sea el precio de las acciones de una empresa, el precio del metro cuadrado de una vivienda en una región, el precio del oro o la plata, o la cantidad de personas que podrían contagiarse de una enfermedad, ha sido de gran importancia para la humanidad. Nos permite estar mejor preparados ante posibles crisis financieras o de salud pública. En este caso, nos enfocaremos en la predicción del número de contagios de COVID-19. Existen diversos métodos para realizar estas predicciones. Algunos de ellos son los modelos SIR (Susceptible-Infectado-Recuperado) y sus variantes, como el modelo SEIR (Susceptible-Expuesto-Infectado-Recuperado). También se utilizan las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN). En nuestro caso, emplearemos modelos basados en series de tiempo, debido a que la inicidencia del virus al igual que en otras pandemias no es constante a lo largo del tiempo. Pueden existir aumentos o disminuciones espontáneas, además de momentos donde la incidencia sea estacionaria. Mediante las series de tiempo es posible identificar tendencias a largo plazo dándonos una aproximación sobre el posible número de contagiados en los meses o incluso años futuros . Es importante destacar que ninguna predicción es infalible y está sujeta a diversas variables y factores.

Para llevar a cabo este proyecto se utilizan datos recolectados por el Instituto Nacional de Salud, disponibles para el público en la página de esta entidad (INS). Con estos datos se establecerá un modelo ARIMA y un modelo GARCH que mejor se ajuste a los mismos, con el fin de describir el comportamiento de la dinámica de infectados por COVID-19 en Santander.

Se espera con este trabajo dar una aproximación del número de casos de contagios de COVID-19 en la región de Santander usando series de tiempo en particular los modelos ARIMA y GARCH.

Planteamiento

Descripción del problema

En el año 2020, la pandemia de COVID-19 fue declarada a nivel mundial, comenzando en China y llegando oficialmente a Colombia el 31 de marzo de 2020 [1]. Siendo un virus desconocido, las autoridades sanitarias carecían de suficiente información sobre su comportamiento y, aún más preocupante, sobre su capacidad de contagio. El COVID-19 no solo ha representado una amenaza directa para la salud de los habitantes, sino que también ha impactado negativamente en el sistema de salud, la economía, la educación y otros sectores vitales.

Justificación

La incertidumbre acerca de cómo se propaga el virus en el territorio colombiano genera pánico, especialmente en cuanto a las medidas sanitarias que deben implementarse y los recursos necesarios para hacer frente a esta nueva pandemia. Por esta razón, resulta imprescindible realizar una aproximación del número potencial de contagios utilizando la información que se recolecta día a día. Se elige implementar un modelo ARIMA, debido a que el número de personas infectadas depende de la cantidad de personas infectadas los días pasados y esté modelo recoge dicha información.

Antecedentes

Un primer trabajo realizado al inicio de la pandemia fue llevado a cabo por ([2]), quien propuso un modelo ARIMA utilizando la metodología de Box-Jenkins. En esta investigación, se abordó la teoría de los modelos de series de tiempo durante un periodo muy temprano de la pandemia, específicamente del 6 de marzo al 28 de octubre de 2020 y del 16 de julio al 28 de octubre del mismo año.

---

[1] Instituto Nacional De Salud. Url: `https://www.ins.gov.co/Noticias/paginas/coronavirus.aspx`. 2020.

[2] Ortiz Cardona. “Propuesta de modelo ARIMA para la serie temporal de los casos de Covid-19 en Colombia aplicando la metodología Box and Jenkins”. En: (2020).

Debido a que en ese momento había disponibles solo un número limitado de datos, el autor llegó a la conclusión de que era necesario realizar una aproximación con un modelo ARIMA de orden muy elevado para tratar de comprender y predecir el comportamiento de la pandemia.

Objetivos

Objetivo general Modelar la dinámica poblacional de la población infectada del COVID-19 en Santander, y su variabilidad, a través de modelos de Series Temporales.

Objetivos específicos

I) Proponer un modelo ARIMA estacional para explicar la dinámica de la población infectada por COVID-19 en Santander.

II) Proponer un modelo con heterocedasticidad (GARCH y/o derivados) para explicar la variabilidad de la dinámica de la población infectada por COVID-19 en Santander.

1. Modelización con modelos ARIMA

1.1. Procesos estocásticos

Dado que para poder definir que es una serie de tiempo necesitamos primero saber que es proceso estocástico se hace necesario definirlo, en los anexos se encuentran las definición que amplian la teoría sobre las variables aleatorias, y teoria de probabilidad.

Definición 1.1.1. [3] Un proceso estocástico es una familia de variables aleatorias $\{X_t\}_{t\in T}$, definidas sobre el espacio de probabilidad $(\Omega, \Im, P)$ y con valores en un espacio medible $(S, \mathfrak{S})$. El conjunto T es conocido como el **conjunto de índices del proceso** y S como el **espacio de estados**.

Definición 1.1.2. [4] Una serie de tiempo es una secuencia ordenada de observaciones cada una de las cuales está asociada a un momento del tiempo, por ejemplo, $t_1, t_2, ..., t_n$.

Definición 1.1.3. [4] Una serie de tiempo univariante consiste en un conjunto de observaciones de una variable Y . Si hay T observaciones, se denota por:

$$Y_t, \qquad t = 1, ..., T.$$

Usualmente t denotará un tiempo. Por ejemplo, considérese $X(t)$ como el casos de COVID-19 en el tiempo t. El proceso $\{X_t\}_{t\geq 0\in T}$ es el número de casos en Santander durante el tiempo $t \geq 0$. Al tomar t fijo y un punto muestral $\omega \in \Omega$, entonces $X_t(\omega)$ se puede escribir también como $X(t, \omega)$.

Ahora vamos a definir que es la función de distribución de un proceso estocastico y un proceso estacionario, dado que como veremos la estacionariedad es fundamental para definir los modelos con series de tiempo.

---

[3] Andrés Sebastián Ríos Gutiérrez. “Modelos epidemiológicos estocásticos y su inferencia: casos SIS y SEIR”. En: *Departamento de Estadística* (2018).

[4] María Pilar González Casimiro. “Análisis de series temporales: Modelos ARIMA”. En: (2009).

Definición 1.1.4. [4] Para conocer la función de distribución de un proceso estocástico es necesario conocer las funciones de distribución univariantes de cada una de las variables aleatorias del proceso, $F[Y_{t_i}], \forall t_i$, y las funciones bivariantes correspondientes, a todo par de variables aleatorias del proceso, $F[Y_{t_i}, Y_{t_j}], \forall (t_i, t_j)$, y todas las funciones trivariantes,...

En resumen, la función de distribución de un proceso estocástico incluye todas las funciones de distribución de un proceso estocástico incluye todas las funciones de distribución para cualquier subconjunto finito de variables aleatorias del proceso:

$$F[Y_{t_1}, Y_{t_2}, ..., Y_{t_n}],\ \forall (t_1, ..., t_n),\ siendo\ n\ finito$$

Definición 1.1.5. [5]Si para $t_1, t_2, ..., t_n$ arbitrarios, tal que $t_1 < t_2 < ... < t_n$, entonces la distribución de las variables aleatorias $(X_{t_1}, X_{t_2}, ..., X_{t_n})$ y $(X_{t_{1+h}}, X_{t_{2+h}}, ..., X_{t_{n+h}})$ son iguales para todo $h > 0$, entonces el proceso estocástico $\{X_t; t \in T\}$ se dice que es un **proceso estacionario de orden** n (o simplemente un proceso estacionario).

Definición 1.1.6. [4] Se define la función de media de un proceso $\{Y_t\}_{t\geq 0}$ sobre la σ-álgebra $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), P)$, como

$$E(Y_t) = \mu_t < \infty, \quad t = 0, \pm 1, \pm 2, ..., \tag{1.1}$$

la función de autocovarianzas se define por.

$$E(Y_t) = E[Y_t - \mu_t]^2 = \sigma_t^2 < \infty \ \ t = 0, \pm 1, \pm 2, ..., \tag{1.2}$$

$$Cov(Y_t, Y_s) = E\left([Y_t - \mu_t][Y_s - \mu_s]\right) = \gamma_{t,s}, \quad \forall t, s (t \neq s) \tag{1.3}$$

Si la distribución del proceso es normal y se conocen sus dos primeros momentos (medias, varianzas y covarianzas), el proceso está perfectamente caracterizado y se conoce su función de distribución.[4]

Definición 1.1.7. [4] Un proceso estocástico Y_t, es estacionario en covarianza(o estacionario en sentido débil) si, y solo si:

---

[5] Liliana Blanco Castañeda, Viswanathan Arunachalam y Selvamuthu Dharmaraja. *Introduction to probability and stochastic processes with applications*. John Wiley & Sons, 2012.

(i) Es estacionario en media, es decir, todas las variables aleatorias del proceso tienen la misma media y es finita:

$$E(Y_t) = \mu < \infty, \quad \forall t > 0 \tag{1.4}$$

(ii) Todas las vairables aleatorias tienen la misma varianza y es finita, es decir, la dispersión en torno a la media constante a lo largo del tiempo es la misma para todas las variables del proceso

$$E(Y_t) = E[Y_t - \mu]^2 = \sigma_Y^2 < \infty, \quad \forall t > 0 \tag{1.5}$$

(iii) Las autocovarianzas sólo dependen del número de periodos de separación entre las variables y no del tiempo, es decir, la covarianza lineal entre dos variables aleatorias del proceso que disten k periodos de tiempo es la misma que existe entre cualesquiera otras dos variables que estén separadas también k peridos, independientemente del momento concreto de tiempo al que estén referidas

$$Cov(Y_t, Y_s) = E[Y_t - \mu][Y_s - \mu] = \gamma_k < \infty, \ \ \forall k \in \mathbb{N} \tag{1.6}$$

Ahora vamos a definir la función de auto-correlación y de auto-correlación parcial, las cuales modela la dependencia de X_t con respecto a $X_{t-1}, X_{t-2}, ...$

Definición 1.1.8. [4]El coeficiente de auto-correlación de orden k de un proceso estocástico estacionario mide el grado de asociación lineal existente entre dos variables aleatorias del proceso separadas k periodos, como sigue:

$$\rho_k = \frac{cov(Y_t Y_{t+k})}{\sqrt{E(Y_t)E(Y_{t+k})}}$$

[4]La función de auto-correlación(FAC) de un proceso estocástico estacionario es una función de k que recoge el conjunto de los coeficientes de auto-correlación del proceso y se denota por $\rho_k, \ k = 0, 1, 2, ...,$. La función de auto-correlación se suele representar gráficamente por medio de un gráfico de barras denominado correlograma(el eje horizontal representa los desfases o retrasos en la serie de tiempo, mientras que el eje vertical muestra los valores de autocorrelación.). Además la función cuenta con las siguientes características:

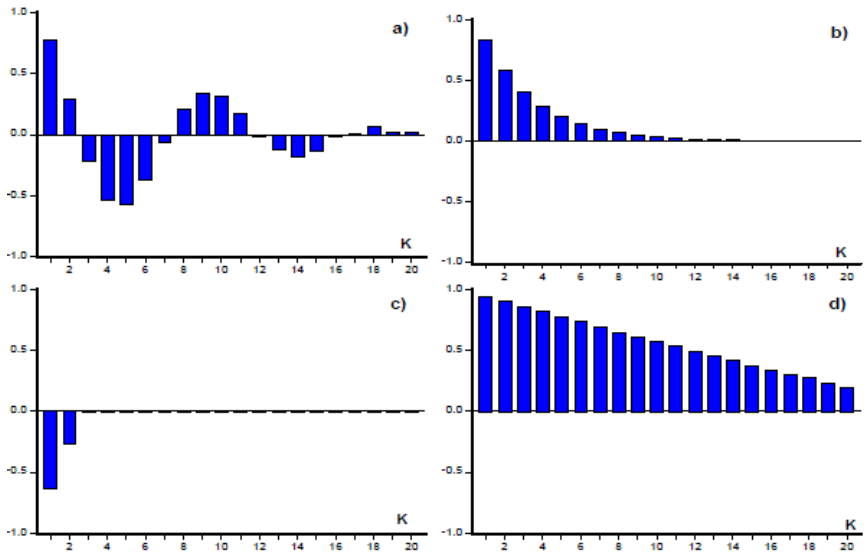


Figura 1.1: Correlograma

- El coeficiente de autocorrelación de orden 0 es, por definición, 1. Por eso, a menudo, no se incluye explícitamente en la función de autocorrelacíon.

$$\rho_0 = \frac{\gamma_0}{\gamma_0} = 1$$

- Es una función simétrica: $\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_{-k}}{\gamma_0} = \rho_{-k}$. Por ello, en el correlograma se representa la función de auto-correlación solamente para los valores positivos del retardo k.

- La función de auto-correlación de una proceso estocástico estacionario tiende a cero rápidamente cuando k tiende a ∞.

La función de autocorrelación va a ser el principal instrumento utilizado para recoger la estructura dinámica lineal del modelo. Los gráficos de la figura (1.1) muestran 4 correlogramas correspondientes a diferentes series temporales. Los correlogramas a), b) y c) decrecen rápidamente hacia cero conforme aumenta k: exponencialmente en los casos a) y b) y truncándose en el caso c). Son, por lo tanto, correlogramas correspondientes a series estacionarias. Por el contrario, los coeficientes de autocorrelación del correlograma d) decrecen lentamente, de forma lineal, por lo que no corresponden una serie estacionaria.

Definición 1.1.9. [4]El coeficiente de auto-correlación parcial de orden k, denotado por p_k, mide el grado de asociación lineal existente entre las variables Y_t e Y_{t-k} una vez ajustado el efecto lineal de todas las variables intermedias, es decir:

$$p_k = \rho(Y_t, Y_{t-k}|Y_{t-k+1}, Y_{t-k+2}, ..., Y_{t-1})$$

Por lo tanto, el coeficiente de auto-correlación parcial p_k es el coeficiente de la siguiente regresión lineal:

$$\tilde{Y}_t = \alpha + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + ... + \rho_k Y_{t-k} + e_t$$

$$p_k = \rho(Y_t, \tilde{Y}_t)$$

Las propiedades de la función de auto-correlación parcial (FACP), $p_k,\ k = 0, 1, 2, ...$, son equivalentes a las de la FAC:

- $p_0 = 1$ y $p_1 = \rho_1$
- Los coeficientes p_k no dependen de unidades y son menores que la unidad en valor absoluto.
- La FACP es una función simétrica
- La FACP de un proceso estocástico estacionario decrece rápidamente hacia cero cuando $k \ \rightarrow \ \infty$.

Una forma intuitiva de entender un proceso que es estacionario en covarianza, es pensar en un proceso que se comporta de manera similar en todos los instantes de tiempo. Por ejemplo, pensemos en el proceso que representa la cantidad de personas de transitan por una avenida cada hora, entonces podemos pensar que este proceso es estacionario en covarianza si la cantidad promedio y la variabilidad de personas que transitan no cambia sin importar la hora del día.

Definición 1.1.10. [4]El ruido blanco es un proceso estocástico con variables aleatorias de media cero, varianza constante y covarianzas nulas. Es deceir es $a_t,\ t = 0, \pm 1, \pm 2, ...$

$$E(a_t) = 0,\ \forall t \quad V(a_t) = \sigma^2,\ \forall t \quad Cov(a_t a_s) = 0,\ \forall t \neq s$$

con $t = 0, 1, ...$

Así, un proceso de ruido blanco, $a_t \sim RB(0, \sigma^2)$, es estacionario si la varianza σ^2 es finita con función de autocovarianzas (FACV):

$$\gamma_k = \sigma^2,\ k = 0\ \ y\ \ \gamma_k = 0,\ k > 0$$

y función de autocorrelación (FAC):

$$\gamma_k = 1,\ k = 0\ \ y\ \ \gamma_k = 0,\ k > 0$$

Definición 1.1.11. [4] Sea Y_t un proceso estocástico esta definido por

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + ... + a_t \quad \forall t = 1, 2, ...$$

Se dice que Y_t es invertible si se cumple que

$$\sum_{i=1}^{\infty} \pi_i^2 < \infty.$$

1.1.1. El teorema de Wold El teorema de descomposición de Wold proporciona una representación de medias móviles de un proceso estocástico estacionario. Además, este teorema establece las condiciones necesarias para la definición del operador de retardo en la modelización de estos procesos.

Teorema 1.1.12. [6] *Supongamos que* $\{X_t\}$ *es un proceso estacionario en covarianza con* $E(X_t) = 0$ *y función de covarianza,* $\gamma(j) = E(X_t X_{t-j})$*. Para todo* $j = 1, 2, ...$ *Entonces*

$$X_t = \sum_{j=0}^{\infty} d_j \epsilon_{t-j} + \eta_t, \tag{1.7}$$

donde,

$$d_0 = 1,\ \sum_{j=0}^{\infty} {d_j}^2 < 0,\ E(\epsilon_t^2) = \sigma_\epsilon^2,\ E(\epsilon_t \epsilon_s) = 0,\ t \neq s, \tag{1.8}$$

---

[6] Christiano Lawrence. *Basic Time Series Analysis. Finance 520-1. General Seminar in Finance*. Inf. téc. Kellogg School of Management, Spring, 2011. (Visitado 2011).

$$E(\epsilon_t) = 0,\ E(\eta_t \epsilon_s) = 0\ \forall t, s, \tag{1.9}$$

$$E[\eta_{t+s}|X_{t-1}, X_{t-2}, ...] = \eta_{t+s}, s \geq 0. \tag{1.10}$$

Definimos el operador de retardo porque es la notación bajo la cual se escribe el proceso ARIMA.

Definición 1.1.13. [7] El operador de retado L se nota como

$$LX_t \equiv X_{t-1}, L^d X_t \equiv X_{t-d}\ \ (d \geq 1) \tag{1.11}$$

[4] Para definir el proceso $ARIMA$, necesitamos antes mencionar el proceso auto regresivo , dicho proceso usa los valores del pasado de la serie para predecir los valores futuros. Además debemos mencionar el proceso medias móviles que nos dará información de la auto-correlación de la serie, también de ayudarnos a identificar patrones en la serie y poder identificar parámetros para el proceso $ARIMA$, el proceso $ARMA$ es la unión de los dos procesos previamente mencionados, así mismo para el proceso $ARIMA$ tenemos que considerar la parte integrada, que es usar una técnica de diferenciación para lograr una serie estacionaria y dar una mejor predicción de la misma.

1.2. Modelos auto rregresivos de media moviles (ARMA)

Definición 1.2.1. [4] El modelo auto regresivo finito de orden p, AR(p) es una aproximación natural al modelo lineal general. Se obtiene un modelo finito:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + a_t \qquad t = 1, 2, ...$$

en términos del operador de retardos se obtiene

$$(1 - \phi L - \phi_2 L^2 - ... - \phi_p L^p) Y_t = a_t \ \ \rightarrow \ \ \phi_p(L) Y_t = a_t$$

---

[7] Peter J Brockwell y Richard A Davis. *Time series: theory and methods*. Springer science & business media, 2009.

donde $\phi_p(L)$ se define como el polinomio auto regresivo.

Enfoque natural, dado que se está considerando que el presente Y_t depende linealmente del ayer Y_{t-1}, el antes de ayer Y_{t-2}, ..., p tiempos atrás Y_{t-p}.

Un proceso auto regresivo es estacionario si las raíces del polinomio auto regresivo con el operador de retardos cae por fuera del circulo unitario, es decir, sus raíces en modulo son mayores a 1. Además un proceso auto regresivo es invertible si tenemos que

$$\sum_{i=1}^{\infty} \phi_i^2 < \infty$$

Definición 1.2.2. [4] El modelo de medidas móviles de orden finito q, MA(q), se define por.

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \qquad a_t \sim RB(0, \sigma^2)$$

en términos del operador de retardos queda como sigue:

$$Y_t = (1 - \theta_1 L - \theta_2 L^2 - ... - \theta_q L^q) a_t \quad \rightarrow \quad Y_t = \theta_q(L) a_t$$

donde $\theta_q(L)$ se define como el polinomio de medias móviles.

Un proceso de medias móviles es invertible si las raíces de su polinomio de medias móviles con el operador de retado cae fuera del circulo unitario. El teorema de Wold garantiza que un proceso estacionario se pueda escribir por medio de medias móviles.

Definición 1.2.3. [4] Los procesos auto regresivos de medias móviles determinan Y_t se define por:

$$Y_t = \phi Y_{t-1} + ... + \phi_p Y_{t-p} + a_t + \theta a_{t-1} + ... + \theta_q a_{t-q} \qquad a_t \sim RB(0, \sigma^2)$$

Este modelo se puede escribir en términos del operador de retados como sigue:

$$(1 - \phi_1 L - ... - \phi_p L^p) Y_t = (1 - \theta_1 L - ... - \theta_q L^q) a_t$$

$$\phi(L) Y_t = \theta_q(L) a_t$$

Teorema 1.2.4. [7] *Si Y_t un proceso ARMA con polinomios $\phi(L)$ y $\theta(L)$ tiene ceros en común entonces, hay dos posibilidades:*
1. Y_t es la única solución estacionaria de las ecuaciones ARMA.
2. La ecuación ARMA puede tener mas de una solución estacionaria.

1.3. Modelos auto rregresivos de medias móviles integrados (ARIMA)

Supongamos el siguiente modelo $ARMA(p,q)$:

$$\Phi_p(L)Y_t = \Theta_q(L)a_t \tag{1.12}$$

donde el polinomio AR se puede factorizar en función de sus p raíces $L_1, ..., L_p$,

$$\Phi_p(L) = (1 - L_1^{-1}L)(1 - L_2^{-1}L)...(1 - L_P^{-1}L)$$

supongamos que $(p-1)$ raíces son estacionarias (con módulo fuera del círculo unidad) y una de ellas es unitaria, $L_i = 1$. Entonces, el polinomio AR, se puede reescribir como sigue:

$$\begin{aligned} &\Phi_p(L) = (1 - L_1^{-1}L)...(1 - L_p^{-1}L) = \varphi_{p-1}(L)(1 - (1)^{-1}L) \\ &\Phi_p(L) = \varphi_{p-1}(L)(1 - L) \end{aligned} \tag{1.13}$$

donde el polinomio $\varphi_{p-1}(L)$ resultada del producto de los $(p-1)$ polinomios de orden 1 asociados a las raíces L_i con módulo fuera del círculo unidad.
Sustituyendo en el modelo $ARMA(p,q)$ (1.12) se tiene que:

$$\varphi_{p-1}(L)(1 - L)Y_t = \Theta_q(L)a_t \;\rightarrow\; \varphi_{p-1}(L)\Delta Y_t = \Theta_q(L)a_t \tag{1.14}$$

donde el polinomio $\varphi_{p-1}(L)$ es estacionario porque todas sus raíces tienen modulo fuera del círculo unidad y el polinomio $\Delta = (1 - L)$ es el que recoge la raíz unitaria.

El modelo (1.14) representa el comportamiento de un proceso Y_t que no es estacionario por que tiene una raíz unitaria. A un proceso Y_t con estas características se le denomina **proceso integrado de orden 1**.

En general, el polinomio AR del modelo (1.12) puede contener más de una raíz unitaria, por ejemplo, d, entonces se puede descomponer como:

$$\Phi_p(L) = \varphi_{p-d}(L)(1-L)^d$$

y sustituyendo, de nuevo, en el modelo $ARMA(p,q)$ (1.12), se tiene:

$$\varphi_{p-d}(L)\Delta^d Y_t = \Theta_q(L)a_t$$

donde el polinomio $\varphi_{p-d}(L)$ es estacionario por que sus $(p-d)$ raíces tienen modulo fuera del círculo unidad, y el polinomio $\Delta^d = (1-L)^d$, de orden d, contiene las d raíces unitarias no estacionarias. A un proceso Y_t con estas características se le denomina **proceso integrado de orden d** y se denota por $Y_t \sim I(d)$.

Definición 1.3.1. [4] Un proceso Y_t es integrado de orden d, $Y_t \sim I(d)$, si Y_t no es estacionario, pero su diferencia de orden d, $\Delta^d Y_t$, sigue un proceso $ARMA(p-d,q)$ estacionario e invertible.

El orden de integración del proceso es el número de diferencias que hay que tomar al proceso para conseguir la estacionariedad en media, o lo que es lo mismo, el número de raíces unitarias del proceso.

Definición 1.3.2. [4] Un modelo auto regresivo integrado de medias móviles si Y_t es integrada de orden d, es de la forma

$$\Phi_p(L)\Delta^d Y_t = \delta + \Theta_q(L)a_t \quad a_t \sim RB(0,\sigma^2) \tag{1.15}$$

donde el polinomio auto regresivo estacionario $\Phi_p(L)$ y el invertible de medias móviles $\Theta_q(L)$ no tienen raíces comunes, y δ es una constante para que su media no sea cero.

Recordemos que gracias al operador de retardos podemos escribir $Y_{t-1} = LY_T$, por tanto

tenemos

$$\begin{aligned}\Delta^{d-1}Y_t - \Delta^{d-1}Y_{t-1} &= (1-L)^{d-1}Y_t - (1-L)^{d-1}LY_t \\ &= (1-L)^{d-1}(Y_t - LY_t) \\ &= (1-L)^{d-1}(1-L)Y_t \\ &= (1-L)^{d}Y_t = \Delta^{d}Y_t.\end{aligned} \tag{1.16}$$

El modelo (1.15) se denomina modelo Autorregresivo Integradro de Medias Móviles de orden (p, d, q) o $ARIMA(p, d, q)$, donde p es el orden del polinomio autorregresivo estacionario, d es el orden de integración de la serie, es decir, el número de diferencias que hay que tomar a la serie para que sea estacionaria, y q es el orden del polinomio de medias móviles invertible.

[4]Muchas series económicas si se observan varias veces a lo largo del a no, trimestral o mensual- mente, presentan comportamiento estacional. Este tipo de comportamiento puede ser debido a factores meteorológicos, tales como la temperatura, pluviosidad, etc.

A la hora de elaborar el modelo ARIMA adecuado para una serie temporal se ha de tener en cuenta el comportamiento estacional, si lo hubiere, porque implica que la observación de un mes y observación del mismo mes del año anterior tienen una pauta de comportamiento similar por lo que estarán temporalmente correlacionadas.Por lo tanto, el modelo de series temporales ARIMA apropiado para este tipo de series debería recoger las dos clases de dependencia intertemporal que presentan, a saber la relación lineal existente entre observaciones sucesivas (comportamiento tendencial o regular) y la relación lineal existente entre observaciones del mismo mes en años sucesivos(comportamiento estacional).

Entonces supongamos que la serie Y_t presenta un componente estacional y se especifica un modelo $ARIMA(p, d, q)$ general:

$$\phi_p(L)Y_t = \theta_q(L)a_t$$

para recoger las dos estructuras de correlación anteriormente mencionadas,la regular y la estacional, bastaría con añadir al modelo los retardos de Y_t y a_t necesarios. Así, si la serie estacional es mensual y se quiere representar, además de la dependencia

entre observaciones consecutivas, la autocorrelación entre observaciones del mismo mes separadas un año, dos años, etc. sería necesario incluir en el modelo retardos hasta de orden 12, 24, etc.

Definición 1.3.3. [4] Se entiende por modelo estacional puro aquel que recoge únicamente relaciones lineales entre observaciones del mismo mes para a nos sucesivos, es decir, entre observaciones separadas s periodos o múltiplos de s, donde $s = 4$ si la serie es trimestral y $s = 12$ si la serie es mensual. Por lo tanto, en estos modelos, para simplificar, se parte del supuesto de que no existe estructura regular, es decir, correlación entre observaciones consecutivas. Como este supuesto es poco rea- lista, este tipo de modelos no va a ser muy útil en la practica pero su estudio permitirá identificar en qué retardos de la función de autocovarianzas y/o autocorrelación se refleja la estructura de tipo estacional de una serie. Se denotarán, en general, $ARMA(P,Q)_s$, donde P es el orden del polinomio autorregresivo estacionario y Q es el orden del polinomio medias móviles invertible.

Veamos ahora los modelos AR y MA estacionales.

Definición 1.3.4. [4] Un proceso de medias moviles estacional $MA(Q)_s$ se define como:

$$\begin{aligned} Y_t &= \Theta_Q(L^s)a_t = (1 - \Theta_1 L^s - \Theta_2 L^{2s} - ... - \Theta_1 L^{Qs})a_t \\ Y_t &= a_t + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + ... + \Theta_Q a_{t-Q_s} \end{aligned} \tag{1.17}$$

Como este proceso es un modelo de medias móviles finito es estacionario para cualquier valor de Θ y su media es cero. El proceso será invertible, si y solo, si las raíces del polinomio de medias móviles tienen modulo fuera del círculo unitario.

Definición 1.3.5. [4] Un proceso autorregresivo estacional $AR(P)_s$ se define como:

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + ... + \Phi_P Y_{t-P_s} + a_t \tag{1.18}$$

Como es un modelo autorregresivo finito es invertible para cualquier valor Φ y su media es cero. El proceso será estacionario, sí y solo, sí las cuatro raíces del polinomio autorregresivo tiene modulo fuera del círculo unidad.

Definición 1.3.6. [4] A partir de las definiciones anteriores podemos definir el modelo

$ARMA(P,Q)_s$ estacional, se define como:

$$\begin{aligned}(1-\Phi_P L^s)Y_t &= (1-\Theta_Q L^s)a_t \\ (1-\Phi_1 L^s-\ldots-\Phi_P L^{P_s})Y_t &= (1-\Theta_1 L^s-\ldots-\Theta_Q L^{Q_s})a_t \\ Y_t &= \Phi_1 Y_{t-s}+\ldots+\Phi_P Y_{t-P_s}+a_t\Theta_1 a_{t-s}+\ldots+\Theta_Q a_{t-Q_s}\end{aligned} \tag{1.19}$$

Este modelo será estacionario cuando las raíces del polinomio autorregresivo tengan módulo fuera del círculo unidad y será invertible cuando las raíces del polinomio de medias móviles tengan módulo fuera del círculo unidad.

[4] De momento hemos supuesto que la serie Y_t es estacionaria, quizás tras algún tipo de transformación. En la práctica, dado que por estacionalidad entendemos una pauta regular de comportamiento cíclico de periodo 1 año de la serie que implica que, en promedio, cada mes tiene un comportamiento diferente, las series estacionales suelen presentar problemas de falta de estacionariedad en media o lo que es lo mismo cambios sistemáticos en el nivel de la serie.

Si la estacionalidad fuese siempre exactamente periódica, $S_t = S_{t-s}$ se podría elimiar de la serie como un componente determinista previamente estimado. Ahora bien, este esquema estacional es muy rígido porque exige que los factores estacionales permanezcan constantes a lo largo del tiempo. Sin embargo, la mayoría de las series no se comportan de una forma tan regular y, en general, el componente estacional será estocástico y estará correlacionado con la tendencia. Por esta razón, un esquema de trabajo más flexible es suponer que la estacionalidad es sólo aproximadamente constante y que evoluciona estocásticamente:

$$Y_t = S_t + \eta_t, \quad S_t = S_{t-s} + \nu_t$$

donde ν_t es un proceso estocástico estacionario.

Para solucionar la no estacionariedad en media que genera el comportamiento estacional, se toman diferencias entre observaciones separadas s periodos, que llamaremos *diferencias estacionales:*

$$\Delta_s Y_t = (1-L^s)Y_t = S_t + \eta_t - S_{t-s} - \eta_{t-s} = \nu_t + \eta_t - \eta_{t-s}$$

de forma que $\Delta_s Y_t \sim MA(1)_s$ es estacionario e invertible.
En conclusión, el operador Δ_s convierte un proceso estacional no estacionario en estacionario.
Un modelo estacional puro se define por.

$$\Phi_P(L^s)\Delta_s^D Y_t = \Theta_Q(L^s)a_t \tag{1.20}$$

En teoría, D, el número de las diferencias estacionales que se han de aplicar para convertir a la serie en estacionaria, puede tomar cualquier valor dependiendo de las características de la serie, pero en la práctica nunca es superior a 1.

[4]) Consideremos una serie estacional Y_t con periodo s conocido, por ejemplo, una serie mensual observada durante N años, de forma que en total contamos con $T = 12N$ observaciones. Si la serie es estacional podemos dividirla en s subseries, una por mes, de N observaciones que denotaremos:

$$y_\tau^{(1)}, y_\tau^{(2)}, y_\tau^{(3)}, ..., y_\tau^{(s)}, \quad \tau = 1, 2, ..., N$$

La relación entre estas subseries y la serie de partida es:

$$y_\tau^{(j)} = Y_{j+s(\tau-1)} = Y_t \quad \tau = 1, 2, ..., N \quad j = 1, 2, ..., 12 \tag{1.21}$$

Cada una de estas doce subseries no presenta comportamiento estacional por lo que podemos representarlas mediante los modelos $ARIMA(p, d, q)$. Supongamos que el modelo $ARIMA$ adecuado para las doce subseries $y_\tau^{(j)}$ es el mismo:

$$(1 - \Phi_1 L - ... - \Phi_p L^p)(1 - L)^D y_\tau^{(j)} = (1 - \Theta_1 L - ... - \Theta_q L^q)u_\tau^{(j)} \quad \tau = 1, 2, ..., N \tag{1.22}$$

Si hay una estacionalidad en la serie de partida Y_t se cumple que $D \geq 1$. Notése que si $D = 0$ y, por lo tanto, las series $y_\tau^{(j)}$ fueran estacionarias, todos los modelos (1.22), al ser el mismo, tendrían la misma media, lo que es incompatible con el supuesto de estacionalidad que implica que la media de cada mes, es decir de cada subserie $y_\tau^{(j)}$, $j = 1, 2, ..., 12$, es diferente.

Los modelos para las doce subseries, al ser todos iguales, se pueden escribir

conjuntamente en función de la serie de partida Y_t, teniendo en cuenta que, dada la relación (1.21):

$$Ly_\tau^{(j)} = y_{\tau-1}^{(j)} = Y_{j+12(\tau-2)} = Y_{j-12+12(\tau-1)} = L^s Y_{j+12(\tau-1)}$$

Lo que implica que aplicar el operador L a $y_\tau^{(j)}$ es equivalente a aplicar el operador L^{12} a la serie original $Y_{j+s(\tau-1)}$.

Además habrá que definir una serie de ruido común para las doce subseries, α_t, asignando a cada mes t el ruido del modelo univariante correspondiente a dicho mes. En consecuencia, α_t se obtendría a partir de las doce subseries $u_\tau^{(j)}$ como sigue:

$$u_\tau^{(j)} = \alpha_{j+12(\tau-1)} \quad \tau = 1, 2, ..., N$$

Aplicando ambos resultado al modelo (1.22), se obtiene el modelo conjunto para toda la serie mensual, Y_t:

$$(1 - \Phi_1 L^{12} - ... - \Phi_p L^{12p})(1 - L^{12})^D y_\tau^{(j)} = (1 - \Theta_1 L^{12} - ... - \Theta_q L^{12q})\alpha_t \quad t = 1, 2, ..., 12N$$

Para cada uno de los modelos (1.22) las series $u_\tau^{(j)}$ son, por construcción, ruido blanco, pero la serie conjunta, $\alpha_1, t = 1, 2, ..., T$, no tiene por qué serlo, en general, ya que la mayoría de las veces existirá dependencia entre las observaciones contiguas que, como no ha sido todavía tenida en cuenta, quedará recogida en esta serie de ruido. Es lo que se denomina *estructura regular* o estructura asociada a los intervalos naturales de medida de la serie.
Suponiendo que α_t sigue el modelo $ARIMA$ no estacional siguiente:

$$\phi_p(L)(1 - L)^d \alpha_t = \theta_q(L)\alpha_t \quad \alpha_t \sim RBN(0, \sigma^2)$$

Sustituyendo este modelo en el general para Y_t se obtiene el modelo completo para la serie observada:

$$\Phi_P(L^s)\phi_p(L)\Delta^d {\Delta_s}^D Y_t = \theta_q(L)\Theta_Q(L^s)\alpha_t \tag{1.23}$$

donde $\phi_p(L)$ y $\theta_q(L)$ son los polinomios autorregresivos y medias móviles de la parte regular, d es el orden de integración de la parte regular y p, q son los ordenes de la parte autorregresiva y de medias móviles estacionarios; mientras que $\Phi_P(L)$ y $\Theta_Q(L)$ son los

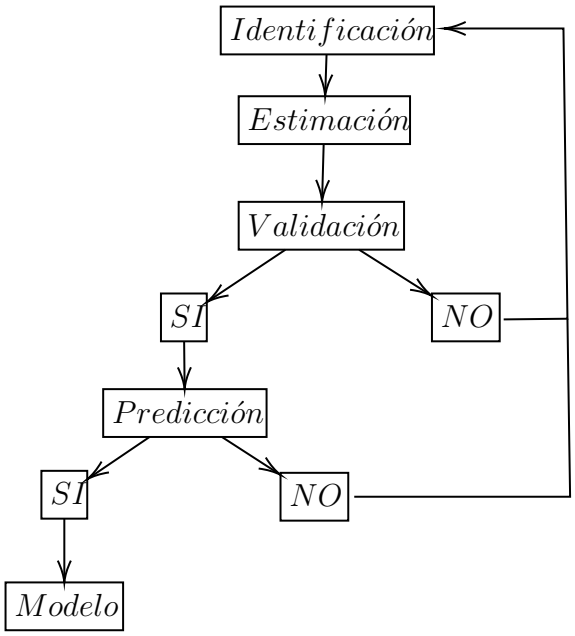
polinomios autorregresivos y medias móviles de la parte estacional, D es el orden de integración de la parte es estacional y P, Q son los ordenes de la parte autorregresiva y de medias móviles estacionales. Este modelo lo llamos **Modelo autorregresivo integrado de medias móviles multiplicativo**.
Los modelos $ARIMA$ estacionales multiplicativos, $ARIMA(p,d,q) \times (P,D,Q)_s$. Estos modelos son flexibles en el sentido de que especifican estacionalidades estocásticas, tendencias estocásticas y además recogen la posible interacción entre ambos componentes.

Esta clase de modelos, como hemos visto, se basa en la hipótesis central de que la relación de dependencia estacional es la misma para todos los periodos. Este supuesto no se tiene por qué cumplir siempre pero, de todas maneras, estos modelos son capaces de representar muchos fenómenos estacionales que encontramos en la práctica de una forma muy simple.

1.4. Modelización con modelos ARIMA

La metodología Box-Jenkins se fundamentan en los siguientes dos principios:

- Selcción de un mode en forma iterativa. En cada etapa se puede presentar la posibilidad de rahacer las etapas previas.



- Principio de parametrización, también denominado parsimonia. Se trata de proponer un modelo capaz de representar la serie con el mínimo de parámetros posibles y únicamente acudir a una amplición del mismo en caso de que sea estrictamente necesario para describir el comportamiento de la serie.

1.4.1. Identificación Con la información de nuestras observaciones y/o de cualquier fuente disponible sobre cómo ha sido generada nuestra serie, se intentará sugerir una subclase de modelos $ARIMA(p,d,q)$ que valga la pena ser estudiada. Nuestro objetivo es encontrar los ordenes p,d,q que parecen apropiados para reproducir las características de la serie bajo estudio y su se incluye o no la constante δ. Es esta etapa cabe la posibilidad de identificar más de un modelo candidato a haber podido generar la serie. Se escoge, el valor de p y q de acuerdo con la función de autocorrelación y parcial, además también se escoge P y Q de acuerdo con la función de autocorrelación. D y d son las diferencias estacionales y no estacionales.

1.4.2. Estimación Utilizando de forma óptima los datos se realiza inferencia sobre los parámetros condicionada a que el modelo investigado sea apropiado.

Dado un determinado proceso definido, se trata de cuantificar los parámetros del mismo, $\theta_1,...,\theta_q,\phi_1,...,\phi_p,\sigma^2$ y , si se necesita la constante δ. Para eso debemos tener un criterio de como encontrar dichos parámetros.

Definición 1.4.1. [8] Sea X una variable aleatoria real de distribución desconocida en la que se extrae una muestra $x_1,...,x_n$ de observaciones independientes. Supóngase también que se dispone de una familia parametriza de funciones de densidad $f_\theta(x)$(es decir, que existe una función de densidad $f_\theta(x)$ para cada valor del parámetro $\theta(x)$).

En este caso, $\theta(x)$ juega el papel de parámetro desconocido y se define la función de

[8] Ronald A Fisher. "On the mathematical foundations of theoretical statistics". En: *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.594-604 (1922), págs. 309-368.

verosimilitud $L(\theta)$ de la siguiente manera:

$$L(\theta) = L(\theta|x_1, ..., x_n) = \prod_i f_\theta(x_i)$$

. Ahora para sea f_0 que corresponde $\theta = \theta 0$, que es el verdadero valor del parámetro. Se desea encontrar el valor $\hat{\theta}$(o estimador) que esté lo más próximo posible al verdadero valor θ_0.
En la práctica, dependiendo de la distribución que generó los datos, se suele utilizar el logaritmo de esta función:

$$\hat{\boldsymbol{\ell}}(\theta|x_1, ..., x_n) = \ln L = \sum_i^n \ln f(x_i|\theta).$$

El método de **máxima verosimilitud** estima θ_0 buscando el valor que θ que maximiza $\ln L$. Este es el llamado **estimador de máxima verosimilitud (MLE)** de θ_0:

$$\hat{\theta}_{mle} = \underbrace{arg\ max}_{\theta \in \Theta} \hat{\ell}(\theta|x_1, ..., x_n).$$

Definición 1.4.2. [9]. Sea θ_0 el conjunto de los parámetros de verdad o los que generan el modelo, y sea θ el conjunto de los parámetros de una aproximación o candidato a modelo. Sea Θ el espacio de parámetros para θ. La discrepancia entre el modelo generado y el candidato a modelo se define como

$$d_n(\theta, \theta_0) = E_0\{-2lnL(\theta|Y_n)\}, \tag{1.24}$$

donde E_0 denota la expectativa bajo el modelo generador, y $L(\theta|Y_n)$ representa la verosimilitud.
Ahora para un conjunto dado de estimaciones de máxima verosimilitud $\hat{\theta_n}$,

$$d_n(\hat{\theta_n}, \theta_0) = E_0\{-2 \ln L(\theta|Y_n)\}_{\theta=\hat{\theta_n}} \tag{1.25}$$

proporcionaría una medida útil de la separación entre el modelo generador y el modelo candidato ajustado. Sin embargo, evaluar 1.25 no es posible, ya que hacerlo requiere el

[9] Joseph E Cavanaugh. “Unifying the derivations for the Akaike and corrected Akaike”. En: (1996).

conocimiento de θ. Sin embargo, Akaike en 1973 notó que $-2lnL(\hat{\theta_n}|Y_n)$ sirve como un estimador sesgado de 1.25, y que el ajuste de sesgo

$$E_0\{E_0\{-2\ln L(\theta|Y_n)\}_{\theta=\theta_n}\} - E_0\{-2\ln L(\hat{\theta_n}|Y_n)\} \tag{1.26}$$

a menudo se puede estimar asintóticamente por el doble de la dimensión de $\hat{\theta_n}$.
Así, si dejamos k representar la dimensión de $\hat{\theta_n}$, entonces, el valor esperado de

$$AIC = -2(\hat{\theta_n}|Y_n) + 2k \tag{1.27}$$

Miremos ahora la definición de nuestro criterio de Akaike, para eso debemos definir nuestra función de máxima verosimilitud

Definición 1.4.3. Sean nuestras innovaciones a_t con $t \in \{1, ..., n\}$ de un proceso $ARIMA(p, q)$ que siguen una distribución normal $a_t \sim Normal(0, \sigma^2)$, ahora tenemos que para una innovación a_{t_i} que sigue una distribución normal,por lo cual escribimos lo siguiente

$$a_{t_i} \sim Normal(0, \sigma^2) a_{t_i} = X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - ... - \phi_p X_{t-p} - a_{t_i} + \theta_1 a_{t-1} + ... + \theta_q a_{t-q} \sim Normal(0, \sigma^2) \tag{1.28}$$

con ϕ y θ siendo los coeficientes de la parte autorregresiva y los de la parte de medias móviles. Así se define la función de verosimilitud como:

$$f(a_1, ..., a_t) = \frac{K}{\sigma^n} \exp\left\{ -\sum_{t=t_1}^{t_n} \frac{(X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} - a_{t_i} + \theta_1 + a_{t-1} + ... + \theta_q a_{t-q})^2}{2\sigma^2} \right\} \tag{1.29}$$

ahora debemos encontrar coeficientes $\hat{\phi_1}, ..., \hat{\theta_q}$ y σ^2 que maximicen a nuestra función de verosimilitud $f(a_1, ..., a_t)$ 1.29 y con dichos valores calculamos nuestro criterio

$$AIC = -2\ln f(a_1, ..., a_t) + 2K \tag{1.30}$$

Definición 1.4.4. [10] La prueba McLeodLi verifica la presencia de heterocedasticidad condicional calculando la prueba de Ljung-Box con los datos al cuadro de un modelo

[10] Kung-Sik Chan y Brian Ripley. *TSA: Time Series Analysis*. 2022. URL: `https://cran.r-project.org/package=TSA`.

arima. Su hipótesis es que el modelo no existe herocesdaticidad condicional entre los retardos.

Definición 1.4.5. [11] La prueba de Shapiro-Wilk se usa para contrastar la normalidad de un conjunto de datos. Se plantea como hipótesis nula que una muestra $x_1, ..., x_n$ proviene de una población normalmente distribuida. Es importante dicha prueba, ya que nos ayuda a ver si los supuestos sobre nuestros datos se cumplen para aplicar el modelo tal como la normalidad en los residuales.

Definición 1.4.6. [12] La prueba de Ljung-Box nos dice si un grupo cualquier de autocorrelaciones de una serie de tiempo son diferentes de cero, su hipotesis nula dice que los datos se distribuyen de forma independiente(es decir, las correlaciones en la población de la que se toma la muestra son 0).

1.4.3. Predicción sobre modelos estacionarios Recordemos el teorema de Wold(1.1.12): $\{X_t\}_{t\geq 0}$ es estacionario entonces se puede escribir de la siguiente.

$$\begin{aligned} X_t &= \sum_{i=0}^{+\infty} \psi_i a_{t-i} \\ \psi_0 &= 1;\ \sum_{i=0}^{+\infty} \psi_i < +\infty;\ a_t \sim RB(0, \sigma^2) \end{aligned} \tag{1.31}$$

De esta forma asumiendo $\{X_t\}_{\geq 0}$ estacionario tenemos que $X_t = \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} +$ ....

Queremos predecir X_{T+l} conociendo $X_T, X_{T-1}, X_{T-2}, ..., X_T(l) = E_T(X_{T+l}) = \psi_0 E(a_{T+l}) + \psi_1 E(a_{T+l-1}) + ...$, definiendo:

$$E_T(a_{T+l}) = \begin{cases} a_{T+j} & \text{para valores conocidos,} \\ 0 & \text{para valores desconocidos} \end{cases}$$

Si se conoce el valor de X_T entonces podemos determinar el valor de $a_T = -f(X_{T-1}, X_{T-2}, ...) + X_T$, si no se conoce X_T entonces la innovación a_T no

---

[11] S Shaphiro y MBJB Wilk. "An analysis of variance test for normality". En: *Biometrika* 52.3 (1965), págs. 591-611.

[12] George EP Box y David A Pierce. "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models". En: *Journal of the American statistical Association* 65.332 (1970), págs. 1509-1526.

está determinada por el conjunto de información por lo que $E_T(a_T) = 0$.

En general las bandas de confianza para una serie de tiempo $\{X_t\}_{t\geq 0}$ corresponde a

$$\begin{aligned}(E(X_{T+l}) \pm qnorm(1-\alpha/2) =_{(1-\alpha/2)} \sqrt{Var(e_T(l))}) \\ e_T(l) = \underbrace{E(X_T(l))}_{valor\ esperado} - \underbrace{X_T}_{dato}\end{aligned} \tag{1.32}$$

donde $E_T(X_{T+l}) = X_{T+l}$ para valores conocidos y $E_T(X_{T+l}) = E(X_{T+l}|X_{T,\ldots})$ para valores desconocidos. Nuestro propósito está en establecer los valores de ψ_i para los cuales:

$$Var(e_T(l)) = \sigma^2 \sum_{i=0}^{l-1} \psi_i^2 \tag{1.33}$$

Ejemplo 1.4.7. Consideremos el modelo $ARIMA(0,1,1)$; es decir el modelo dado por:

$$\begin{aligned}(1-L)X_t &= (1+\theta L)a_t \\ X_t &= X_{t-1} + a_t + \theta a_{t-1}\end{aligned} \tag{1.34}$$

La función de predicción está dada por:

$$\begin{aligned}E_T(X_{T+1}) &= E(X_T + a_{T+1} + \theta a_T) = X_T + \theta a_T \\ E_T(X_{T+2}) &= E(X_{T+1} + a_{T+2} + \theta a_{T+1}) = E(X_{T+1}) = X_T + \theta a_T \\ &\vdots \\ E_T(X_{T+l}) &= E(X_{T+l} + a_{T+l} + \theta a_{T+l-1}) = E(X_{T+l}) \\ &= X_T + \theta a_T \ \ para\ todo\ l.\end{aligned} \tag{1.35}$$

De esta manera la banda de confianza es un tubo horizontal dependientes fundamentalmente de la última observación.

En caso de que $\theta = 0$ (es decir, cuando se toma un paseo aleatorio) se tiene entonces que la función de predicción está dada por $E_T(X_{T+l}) = X_T$.

En general se puede demostrar que para un modelo con orden de integración de orden 1 (es decir, por ejemplo un modelo $ARIMA(p,1,q)$) con término independiente ($\delta = 0$) la

función de predicción cuando $l \to +\infty$, tiende a ser constante:

$$X_T(l) = E(X_{T+l}) \xrightarrow[l\to+\infty]{} K^T \tag{1.36}$$

K^T no es la media de todo el proceso (en virtud de que por ser no estacionario el proceso puede seguir una tendencia) K^T es entonces una constante dependiente del conjunto de la información y de los parámetros AR Y MA del modelo no estacionario.

Cuando el modelo tiene termina independiente $\theta \neq 0$ entonces la función de predicción está dada por:

$$\begin{aligned}
E_T(X_{T+1}) &= E(\delta + X_T + a_{T+1} + \theta a_T) = \delta + X_T + \theta a_T \\
E_T(X_{T+2}) &= E(\theta + X_{T+1} + a_{T+2} + \theta a_{T+1}) = \theta + E_T(X_{T+1}) \\
&= 2\delta + X_T + \theta a_T \\
&\vdots \\
E(X_{T+l}) &= E(\delta + X_{T+l-1} + a_{T+l} + \theta a_{T+l-1}) = \delta + E_T(X_{T+l-1}) \\
&= \delta l + X_T + \theta a_T
\end{aligned} \tag{1.37}$$

Razonando de manera análoga a como lo hicimos previamente tenemos que

$$E(X_{T+l}) \xrightarrow[l\to+\infty]{} K^T + \delta l \tag{1.38}$$

Cuando el proceso no es estacionario el limite $\lim_{l\to+\infty} Var(e_T(l))$ no necesariamente exite.

Para establecer $Var(e_T(l))$ seguimos el mismo razonamiento realizado para modelos estacionarios, es decir, determinando ψ_i.

Escribiendo el modelo $ARIMA(p, d, q)$ en representación $MA(\infty)$(entendiendo que no todo modelo MA es estacionario). En el caso del paseo aleatorio se puede escribir como:

$$X_t = a_t + a_{t-1} + a_{t-2} + ... \tag{1.39}$$

de forma que $\psi_i = 1$ para todo $i = 1, 2, 3, ...$ Así

$$
\begin{aligned}
Var(e_T(1)) &= \sigma^2 \\
Var(e_T(2)) &= \sigma^2 \sum_{i=0}^{1} \psi_i^2 = \sigma^2(1+1) = 2\sigma^2 \\
&\vdots \\
Var(e_T(l)) &= \sigma^2 \sum_{i=0}^{l-1} \psi_j^2 = \sigma^2(1+...+1) = l\sigma^2
\end{aligned}
\tag{1.40}
$$

Note que cuando $l \to +\infty$ el correspondiente límite tiende a infinito.

Ejemplo 1.4.8. Modelo $ARIMA(p,2,q)$
La su función de predicción corresponde a $E_T(l) \xrightarrow[l\to+\infty]{} K_1^T + K_2^T l$ donde K_1^T y K_2^T depende del conjunto de información y de los parámetros del modelo. Al tener dos términos, se observa que la predicción de los modelos $ARIMA(p,2,q)$ es más flexible con respecto a $ARIMA(p,1,q)$.

1.4.4. Sobre los errores de predicción $e_T(l) = X_{T+l} - E_T(X_{T+l})$, este es el residual bajo el enfoque de regresión. Veamos como sería el desarrollo.

$$
\begin{aligned}
e_T(1) &= X_{T+1} - E_T(X_{T+1}) \\
&= a_{T+1} + \psi_1 a_T + \psi_2 a_{T-1} + ... - (E_T(a_{T+1}) + \psi_1 E_T(a_T) + \psi_2 E_T(a_{T-1}) + ...) \\
&= a_{T+1} \\
e_T(2) &= X_{T+2} - E_T(X_{T+2}) \\
&= a_{T+2} + \psi_1 a_{T+1} + \psi_2 a_T + ... - (E_T(a_{T+2}) + \psi_1 E_T(a_{T+1}) + \psi_2 E_T(a_T) + ..) \\
&= a_{T+2} + \psi_1 a_{T+1} \\
&\vdots \\
e_T(l) &= X_{T+l} - E_T(X_{T+l}) \\
&= a_{T+l} + \psi_1 a_{T+l-1} + \psi_2 a_{T+l-2} + ... - (E_T(a_{T+l}) + \psi_1 E_T(a_{T+l-1}) + \psi_2 E_T(a_{T+l-2}) + ... \\
&+ \psi_{l-1} E_T(a_{T+1}) + \psi_l E_T(a_T + ...) \\
&= a_{T+l} + \psi_1 a_{T+l-1} + ... + \psi_{l-1} a_{T+1}.
\end{aligned}
\tag{1.41}
$$

$e_T(l)$ es una variable aleatoria con $E(e_T(l)) = 0$ y con varianza dada por:

$$\begin{aligned} Var(e_T(1)) &= E_T((e_T(1))^2) = \sigma^2 \\ Var(e_T(2)) &= E_T((e_T(2))^2) = E_T((a_{T+2} + \psi_1 a_{T+1})^2) \\ &= E_T(a_{T+2}^2 + 2\psi_1 a_{T+2} a_{T+1} + \psi_1^2 a_{T+1}^2) \\ &= \sigma^2 + \psi_1^2 \sigma^2 = (1 + \psi_1^2)\sigma^2 \end{aligned} \tag{1.42}$$

Razonando recursivamente (de hecho es algo que podemos demostrar por inducción) se tiene que :

$$Var(e_T(l)) = (1 + \psi_1^2 + \psi_2^2 + ... + \psi_{l-1}^2)\sigma^2 = \sigma^2 \sum_{i=1}^{l-1} \psi_i^2 \tag{1.43}$$

Si tenemos un modelo pura el cual queremos predecir para $l \to +\infty$. De acuerdo con la invertibilidad nosotros establecimos que:

$$\sum_{i=0}^{+\infty} \psi_i^2 < \infty \tag{1.44}$$

lo que garantiza que $Var(e_T(l))$ cuando $l \to +\infty$ va a converger. Bajo el supuesto de normalidad tenemos que : $e_T(l) = X_{T+l} - E_T(X_T + l) \sim N(0, Var(e_T(L)))$, luego intervalo de confianza corresponde a $(X_T(l)) \pm Z_{(1-\alpha/2)}\sigma\sqrt{\sum_{i=0}^{l-1} \psi_i}$

1.4.5. Predicción con modelos ARMA(p,q) Dado que tenemos la información hasta Y_T vamos a establecer $Y_{T+1}, Y_{T+2}, ...$ para $ARMA(1, 2)$

$$\begin{aligned} Y_{T+1} &= \phi Y_T + a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1} \\ Y_T(1) &= E(Y_{T+1}) = E_T(\phi Y_T + a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1}) \\ &= \theta Y_T - \theta_1 a_T - \theta_2 a_{T-1} \\ Y_{T+2} &= \theta Y_{T+1} + a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T \\ Y_T(2) &= E(Y_{T+2}) = E_T(\phi Y_{T+1} + a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T) \\ Y_{T+3} &= \phi_{T+2} + a_{T+3} - \phi_1 a_{T+2} - \theta_2 a_{T+1} \\ &= \phi E(Y_{T+2}) \end{aligned} \tag{1.45}$$

En general $Y_T(k) = E_T(Y_{T+k}) = \phi E(Y_{T+k})$, para obtener la varianza del error de predicción

vamos a usar que:

$$\begin{aligned}
(1-\phi L)Y_t &= (1-\theta_1 L-\theta_2 L^2)a_t \\
Y_t &= \left(\frac{1-\theta_1 L-\theta_2 L^2}{1-\phi L}\right)a_t = (1+\psi_1 L+...)a_T \\
1-\theta_1 L-\theta_2 L^2 &= (1-\phi L)(1+\psi_1 L+\psi_2 L^2+...) \\
L: \quad -\theta_1 L &= (\psi_1-\theta)L \Rightarrow \psi_1 = \phi-\theta_1 \\
L^2: \quad -\theta_2 L^2 &= (\psi_2-\phi\psi_1)L^2 \Rightarrow \psi_2 = \phi(\phi-\theta_1)-\theta_2 \\
L^3: \quad -\theta_3 L^3 &= (\psi_3-\psi_2\phi)L^3 \Rightarrow \psi_3 = \phi\psi_2
\end{aligned} \tag{1.46}$$

Los pesos de la forma de medias móviles corresponde a :

$$\psi_i = \begin{cases} k=0 & ,\psi_0 = 1 \\ k=1 & ,\psi_1 = \phi-\theta_1 \\ k=2 & ,\psi_2 = \phi\psi_1-\theta_2 \\ k>2 & ,\psi_k = \phi\psi_{k-1} \end{cases} \tag{1.47}$$

De esta manera la varianza del error está dada por:

$$\begin{aligned}
Var(e_t(1)) &= \sigma^2 \\
Var(e_T(2)) &= \sigma^2(1+\phi-\theta_1) \\
Var(e_T(3)) &= \sigma^2(1+\phi-\theta_1+\phi(\phi-\theta_1)-\theta_2) \\
Var(e_T(k)) &= \sigma^2(1+\phi-\theta_1+...+\phi\psi_{k-1})
\end{aligned} \tag{1.48}$$

El intervalo de confianza para $Y_T(l)$ está dada por:

$$\left(Y_T(l) \pm Z_{(1-\alpha/2)}\sqrt{\sum_{i=0}^{l-1}\psi_i}\right) \tag{1.49}$$

2. Modelización con modelos GARCH

Un modelo $ARIMA$ es un modelo homocedástico, es decir, un proceso con igual varianza para toda variable aleatoria definida para cada tiempo. Sin embargo, en busca de explicar la variabilidad o dispersión entre el número de casos reportados entre un día y otro, es que se utilizan los modelos $GARCH$. Para un modelo $GARCH$ es fundamental utilizar el mejor modelo $ARMA$ para el conjunto de datos de acuerdo con el paquete en el lenguaje R. Esto conlleva a establecer una metodología de modelización bajo modelos $ARIMA$ y, luego modelar la variabilidad con modelo $GARCH$.Además se determinan modelos $sGARCH$ y $eGARCH$, siendo el modelo estándar y el modelo exponencial respectivamente.

2.1. Modelo autorregresivo con heterocedasticidad condicional

Proposición 2.1.1. [13] *Dadas dos σ-álgebras $\Im_1, \Im_2$, con $\Im_1 \subseteq \Im_2$ y una variable aleatoria escalar Y, se tiene que:*

$$E(Y \mid \Im_1) = E\left[\, E\,(Y \mid \Im_2) \middle| \, \Im_1\right]$$

Las sigma-álgebras son las generadas por la historia pasada de las variables del modelo, en dos instantes distintos de tiempo. Un caso particular de esta ley que resulta especialmente útil es cuando $\Im_1 = \phi$, pues entonces,

$$E(Y) = E[E(Y|\Im_2)]$$

que relaciona un momento incondicional y un momento condicional.

Sea $\varepsilon_t(\theta)$ denota un proceso estocástico que denota el error, es decir, la distancia entre el pronóstico $(E_{t-1}(X_t))$ y el dato X_t en t. De tiempo discreto con media condicional y funciones de varianza parametrizadas por el vector de dimensión finita $\theta \subseteq \mathbb{R}$, donde θ_0 denota los valores reales. Por simplicidad de notación supondremos inicialmente que $\varepsilon_t(\theta)$ es un escalar, también que $E_{t-1}(\cdot)$ denota la esperanza matemática, condicionada

[13] Alfonso Novales. "Modelos ARCH univariantes y multivariantes". En: *Departamento de Economıa Cuantitativa. Universidad Complutense de Madrid.(Versión Preliminar). Madrid, Espana* (2013).

a la pasada, del proceso, junto con cualquier otra información disponible en el momento $t-1$.

Definición 2.1.2. [14] Decimos que $\varepsilon_t(\theta)$ sigue un proceso ARCH si su esperanza condicional es igual a cero:

$$E_{t-1}(\varepsilon_t(\theta_0)) = 0, \quad t = 1, 2, 3, ...$$

y su varianza condicional,

$$\sigma_t^2(\theta_0) \equiv Var_{t-1}(\varepsilon_t(\theta_0) = E_{t-1}(\varepsilon_t^2(\theta_0))$$

depende, en forma no trivial, del σ-álgebra $\Im_{t-1}$ generada por las observaciones pasadas. La notación σ_t^2 hace referencia al hecho de que trabajamos con un segundo momento del proceso estocástico. Debe apreciarse que, a pesar del subíndice temporal, σ_t^2 es una función de variables pertenecientes al instante $t-1$ o anteriores.

La esperanza y varianza incondicionales del proceso $\varepsilon_t(\theta_0)$ son la esperanza matemática de los momentos análogos condicionales,

$$E(\varepsilon_t) = E(E_{t-1}\varepsilon) = 0 \tag{2.1}$$

$$Var(\varepsilon_t) = E(\varepsilon_t^2) = E(E_{t-1}(\varepsilon_t^2)) = E(\sigma_t^2) = +\alpha(L)\varepsilon_{t-1}^2 \tag{2.2}$$

Su estructura basica es la siguiente:

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t, \quad z_t \sim RB(0,1) \\ \sigma_t^2 &= \omega + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2, \ \alpha_0 > 0, \ \alpha_i \geq 0, \ \sum_{j=1}^{q} \alpha_i \geq 1 \end{aligned} \tag{2.3}$$

Una forma intuitiva de entender el modelo $ARCH$ al igual que los modelos AR (1.2.1) es entender que el valor de la volatilidad en un tiempo t depende de un ruido blanco y

[14] Tim Bollerslev, Robert F Engle y Daniel B Nelson. "ARCH models". En: *Handbook of econometrics* 4 (1994), págs. 2959-3038.

del valor de su varianza en un tiempo t. Dicha varianza depende de la volatilidad en un tiempo $t-1$, aunque el modela $ARCH$ nos ayuda para modelar la volatilidad dependiendo de sus valores pasados. Generalmente suele tener un orden elevado además de que la volatilidad en cada momento explica la de cada momento siguiente por eso, definimos un modelo $ARCH$ generalizado,

2.1.1. Modelo ARCH generalizado (GARCH)

Definición 2.1.3. [14] . El **modelo** $GARCH(p,q)$ es,

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \equiv \omega + \alpha(L)\varepsilon_{t-1}^2 + \beta(L)\sigma_{t-1}^2. \tag{2.4}$$

donde σ_t^2 denota la varianza condicional, ε_t^2 las innovaciones, α y β son coeficientes de la parte autorregresiva y los coeficientes que determinan el impacto de los errores pasados para la volatilidad actual, y ω es el intercepto. Este modelo también es conocido como el **modelo GARCH estandar** (sGARCH).

Para que la varianza condicional en el modelo $GARCH(p,q)$ esté bien definida, todos los coeficientes en el modelo $ARCH$ lineal de orden infinito correspondiente deben ser positivos. Siempre que $\alpha(L)$ y $\beta(L)$ no tenga raíces comunes y que las raíces de los polinomios $\beta(x) = 1$ se encuentren fuera del círculo unitario, esta restricción de positividad se cumple, si y solo, si todos los coeficientes en la expansión de la serie de potencias infinitas $\alpha(x)/(1-\beta(x))$ son no negativas.

Los modelos $GARCH$ recogen adecuadamente las propiedades de agrupamiento de la volatilidad, pero son simétricos, ya que la varianza condicional depende de la magnitud de las innovaciones retardadas, pero no de su signo. Es decir el impacto de las innovaciones no distingue la diferencia si es positiva o negativa. Para recoger los efectos de mas innovaciones no símetricas se propone el modelo exponencial $GARCH$, o $EGARCH(p,q)$.

Definición 2.1.4. [15] El modelo $EGARCH(p,q)$ se define:

$$\ln(\sigma_t^2) = \omega + \sum_{j=1}^{q}(\alpha_j z_{t-j} + \gamma_j(|z_{t-j}| - E|z_{t-j}|)) + \sum_{j=1}^{p}\beta_j \log_e(\sigma_{t-j}^2) \tag{2.5}$$

donde los coeficientes α_j capturan los efectos del signo y γ_j el tamaño del efecto.

De este modo, $\{\ln(\sigma_t^2)\}$ sigue un proceso $ARMA(p,q)$, con las condiciones habituales de estacionariedad del modelo $ARMA$ usual. Como en el caso del $GARCH$, ω fácilmente se puede hacer una función del tiempo para acomodar el efecto de cualquier evento predecible.

2.2. Modelización con modelos GARCH

2.2.1. Estimación [13] La estimación se lleva a cabo, generalmente, por máxima verosimilitud, para lo que suponemos una determinada densidad $f(w_t, \eta)$ para el término de error tipificado,

$$w_t(\theta) = \frac{z_t(\theta)}{\sigma_t} = \frac{\varepsilon_t - \mu_t(\theta)}{[\sigma_t^2(\theta)]^{1/2}}$$

que tiene esperanza cero y varianza uno. Dado un vector de observaciones $\{\varepsilon_1, ..., \varepsilon_T\}$, el logaritmo de la función de verosimilitud para la observación t es:

$$\ell_t(\varepsilon_t; \mu) = \ln\left\{f(w_t(\theta), \mu) - \frac{1}{2}\ln(\sigma_t^2(\theta))\right\}$$

donde el último término es el Jacobiano de la transformación que pasa de las innovaciones estandarizadas a las observaciones muestrales, que en el caso multivariante se convertirá en:

$$\ell_t(\varepsilon_t; \mu) = \ln\left[f(z_t(\theta)[\Gamma_t(\theta)]^{-1}; \mu)\right] - \frac{1}{2}\ln|\Sigma_t(\theta|$$

donde Γ es una matriz no singular(matriz cuadrada con inversa), de igual dimensión que Σ, tal que $\Gamma\Gamma' = \Sigma$(recodando que Γ' es la matriz transpuesta). Se sabe que para toda matriz de definida positiva Σ existe tal matriz Γ. Si la matriz Σ es diagonal, aunque con

---

[15] Alexios Ghalanos. "Introduction to the rugarch package.(Version 1.3-1)". En: *Manuscript, http://cran. r-project. org/web/packages/rugarch. Accessed* 11 (2020).

elementos diferentes a lo largo de la diagonal principal, entonces Γ es la matriz diagonal que tiene por elementos la raíz cuadrada de los elementos en la diagonal de Σ. Como los elementos de esta última, los $\sigma_t^2(\theta)$ son todos positivos.

Por otra parte, utilizando un argumento estándar para la descomposición del error de predicción, la función de verosimilitud para la muestra completa puede escribirse como la suma de los logaritmos de la función de verosimilitud condicional:

$$L_T(\varepsilon_1, ..., \varepsilon_T) = \sum_{t=1}^{T} \ell_t(\varepsilon_t; \psi)$$

cuya maximamización generará estimadores de máxima verosimilitud del modelo, $\psi = (\theta, \mu)$.

2.2.2. Predicción Debido a una ausencia de convergencia en la función de verosimilitud se opto por un método usando bootstrap para la predicción. Para la predicción del σ_t, es decir la varianza y los posible valores de la serie, usaremos la metodología explicada en [16], para un modelo $GARCH(1,1)$ pero vale la generalización de la misma forma para cualquier se hace la deducción modelo $GARCH(p,q)$.

El modelo $GARCH(1,1)$ esta dado por

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 \quad t = 1, ..., T. \end{aligned} \tag{2.6}$$

Donde ε_t es un proceso de ruido blanco con varianza 1, σ_t es un proceso estocástico que es conocido como la volatilidad (o la varianza) y asumamos que son independientes de ε, y ω, α y β son parámetros desconocidos que satisfacen $\omega > 0$, $\alpha \geq 0$,$\beta \geq 0$ para que la varianza condicional sea positiva. El proceso y_t es estacionario en covarianza si, $\alpha + \beta < 1$. Se puede mostrar que y_t es estrictamente estacionario si $E[\log(\beta + \alpha\varepsilon_t^2)] < 1$.

Note que σ_t^2 es observable con la información $t-1$ y en consecuencia, dados los

---

[16] Lorenzo Pascual, Juan Romo y Esther Ruiz. "Bootstrap prediction for returns and volatilities in GARCH models". En: *Computational Statistics & Data Analysis* 50.9 (2006), págs. 2293-2312.

supuestos sobre la distribución de ε_t, la media condicional de y_t es cero y σ_t^2 es la varianza condicional, además, la distribución condicional de y_t coincide con la distribución de ε_t. La varianza condicional se puede ver como, (dicha deducción para la varianza se puede ver [17]).

$$\sigma_t^2 = \frac{\omega}{1-\alpha-\beta} + \alpha \sum_{j=0}^{\infty} \beta^j \left(y_{t-j-1}^2 - \frac{\omega}{1-\alpha-\beta} \right). \tag{2.7}$$

Definición 2.2.1. [18] El error cuadrático medio (ECM) de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. El ECM es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa.
El ECM de un estimador $\hat{\theta}$ con respecto al parámetro desconocido θ se define como

$$ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

El predictor de y_{T+k}, dadas las obervaciones del proceso hasta el momento T, $Y_t = \{y_1, ..., y_T\}$, es cero y su error cuadrático medio condicional (MSE) es dado por

$$E_T(\sigma_{T+k}^2) = E(\sigma_{T+k}^2 | Y_0, Y_T) = \frac{\omega}{1-\alpha-\beta} + (\alpha+\beta)^{k-1} \left(\sigma_{T+1}^2 - \frac{\omega}{1-\alpha-\beta} \right) \tag{2.8}$$

Suponiendo que ε_t es un proceso $\varepsilon_t \sim Normal(0,1)$, entonces y_t es condicionalmente gaussiano(es decir sigue una distribución normal condicional, por tanto la distribución de y_t depende de la información previa) y los errores de pronosticó de a un paso adelante se distribuyen normalmente, por tanto el intervalo de confianza para Y_{T+1} está dado por $(E_T(\varepsilon_{T+1}) \pm qnorm(1-\alpha/2)E(\sigma_{T+1}))$

Por el contrario, la distribución del error en la predicción de k-periodos adelante con $k > 1$ no es normal incluso si ε_t es Gaussiano, la aproximación habitual del intervalo de predicción $(1-\alpha)\,\%$ para los retornos y_{T+k} para $k > 1$ es dado por $(E_T(\varepsilon_{T+k}) \pm qnorm(1-\alpha/2)E(\sigma_{T+k}))$. Los puntos para la predicción del σ_{T+k}^2 esta dados

---

[17] Brandon Williams. “Betreuung: Prof. Dr. Rainer Dahlhaus”. En: (2011).

[18] Erich L Lehmann y George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

por (2.8).

Ahora sea Y_t una sucesión de observaciones generados por $GARCH(1,1)$ dado por (2.6). El objetivo es estimar directamente la distribución de y_{T+k} y σ_{T-k} condicional a los datos disponibles. Los parámetros del modelo, $\theta = (\omega, \alpha, \beta)$, son estimados por $\hat{\theta}_T = (\hat{\omega}, \hat{\alpha}, \hat{\beta})$. Los residuales se computan por $\hat{\varepsilon}_t = y_t/\hat{\sigma}_t$, $t = 1, ..., T$ donde $\hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha} y_{t-1}^2 + \hat{\beta}\hat{\sigma}_{t-1}^2$, con $t = 2, ..., T$, son las varianzas condicionales estimadas y $\hat{\sigma}_1^2 = \hat{\omega}/(1 - \hat{\alpha} - \hat{\beta})$ es la estimación de la varianza marginal.

Para implementar la técnica bootstrap, es necesario obtener replicas bootstrap $Y_t^* = \{y_1^*, ..., y_T^*\}$, que imitan la estructura de la serie original, Estas replicas se obtiene de las siguientes formulas recursivas:

$$\hat{\sigma}_t^{*2} = \hat{\omega} + \hat{\alpha} y_{t-1}^{*2} + \hat{\beta}\hat{\sigma}_{t-1}^{*2} \tag{2.9}$$

$$y_t^* = \varepsilon_t^* \hat{\sigma}_t^*, \quad t = 1, ..., T, \tag{2.10}$$

donde ε_t^* son sorteos aleatorios con reemplazo de $\hat{F}_T$, la función de distribución empírica de los residuos centrados, y $\hat{\sigma}_1^{*2} = \hat{\sigma}_1^2$. Una vez los parámetros de esta serie bootstrap son estimados, $\hat{\theta}_T^* = (\hat{\omega}^*, \hat{\alpha}^*, \hat{\beta}^*)$, la predicción bootstrap de los valores futuros se obtiene de las siguientes formulas recursivas:

$$\hat{\sigma}_{T+k}^{*2} = \hat{\omega}^* + \hat{\alpha}^* y_{T+k-1}^{*2} + \hat{\beta}^* \hat{\sigma}_{T+k-1}^{*2} \tag{2.11}$$

$$y_{T+k}^* = \varepsilon_{\hat{T+k}}{}^* \hat{\sigma}_{T+k}^*, \quad k = 1, 2, ... \tag{2.12}$$

Con ε_{T+k}^* siendo sorteos aleatorios con reemplazos de $\hat{F}_T$,$y_T^* = y_T$ y

$$\hat{\sigma}_T^{*2} = \frac{\hat{\omega}^*}{1 - \hat{\alpha}^* - \hat{\beta}^*} + \hat{\alpha}^* \sum_{j=0}^{T-2} \hat{\beta}^{*j} \left(y_{T-j-1}^2 - \frac{\hat{\omega}^*}{1 - \hat{\alpha}^* - \hat{\beta}^*} \right) \tag{2.13}$$

Se puede apreciar que en la expresion (2.13), aunque $\hat{\sigma}_T^{*2}$ es diferente para todas las replicas del bootstrap, su valor se obtiene utilizando las estimaciones de los parámetros bootstrap correspondientes y siempre la serie original. Por tanto, su valor es pequeño cuando los rendimientos al final del período de la muestra son pequeños y grande

cuando son grandes en valor absoluto. Por tanto $\hat{\sigma}_T^{*2}$ incorpora la variabilidad debida a la estimación de parámetros y, al mismo tiempo, tiene en cuenta el estado del proceso cuando se realizan las predicciones.

Una vez tenemos el conjunto de réplicas de arranque B, $\left(y_{T+k}^{*(1)}, ..., y_{T+k}^{*(B)}\right)$ para y_{T+k}, los límites de predicción se definen como los cuantiles de la función de distribución de arranque de y_{T+k}^*. Mas específicamente, si $G_y^*(h) = Pr(y_{T+k}^* \leq h)$ es la función de distribución de y_{T+k}^* y su estimación es $G_{y,B}^*(h) = \#(y_{T+k}^{*b} \leq h)/B$, donde $\#(\cdot)$ cuenta el numero de casos donde la condición que esta entre paréntesis se satisface, entonces, el intervalo de predicción $100(1-\alpha)\,\%$ para y_{T+k}^* esta dado por

$$\left[L_{y,B}^*(y), U_{y,B}^*(y)\right] = \left[Q_{y,B}^*\left(\frac{\alpha}{2}\right), Q_{y,B}^*(1-\frac{\alpha}{2})\right], \tag{2.14}$$

donde $Q_{y,B}^* = G_{y,B}^{*-1}$.
Podemos simultáneamente obtener la perdición del intervalo de volatilidad para el perdió k futuro. Dado el conjunto $\left(\sigma_{T+k}^{*(1)}, ..., \sigma_{T+k}^{*(B)}\right)$ de B réplicas de arranqué de la volatilidad para cualquier horizonte k, se procede de la misma forma que anteriormente se menciono, utilizando como límites de predicción los cuantiles de la función de distribución bootstrap de $\hat{\sigma}_{T+k}^*$. En este caso, si $G_{\sigma,B}^*(h) = Pr(\hat{\sigma}_{T+k}^* \leq h)$ es la distribución de la función de $\hat{\sigma}_{T+k}^*$ y su estimación de Monte Carlo es $G_{\sigma,B}^*(h) = \#(\hat{\sigma}_{T+k}^{*b} \leq h)/B$, el intervalo de predicción $100(1-\alpha)\,\%$ para $\hat{\sigma_{T+k}^*}$ esta dado por

$$\left[L_{\sigma,B}^*(\sigma), U_{\sigma,B}^*(\sigma)\right] = \left[Q_{\sigma,B}^*\left(\frac{\alpha}{2}\right), Q_{y,B}^*\left(1-\frac{\alpha}{2}\right)\right], \tag{2.15}$$

donde $Q_{\sigma,B}^* = G_{\sigma,B}^{*-1}$.

2.2.3. Estimación [7] Para la estimación del modelo $GARCH(p,q)$ usaremos el criterio de Akaike al igual que la estimación del modelo $ARIMA$. Sean $\{\widetilde{\varepsilon}\}_t$ las (posibles) observaciones de los errores (innovaciones). Los coeficientes GARCH se estiman maximizando numéricamente la probabilidad de $\widetilde{\varepsilon}_{p+1}, \ldots, \widetilde{\varepsilon}_n$ condicional a los valores conocidos de $\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_p$, y asumimos que el valor es 0 para los $\widetilde{\varepsilon}_t$, $t \leq 0$, y $\hat{\sigma}^2$ para cada valor de σ_t^2, donde $\hat{\sigma}^2$ es la varianza muéstrala de $\{\widetilde{\varepsilon}_1, ..., \widetilde{\varepsilon}_n\}$. Donde buscamos maximizar

la siguiente función de verosimilitud.

$$L(\omega, \alpha_1, ..., \alpha_q, \beta_1, ..., \beta_p) = \prod_{t=p+1}^{n} \frac{1}{\sigma_t} \phi \left(\frac{\widetilde{\varepsilon}_t}{\sigma_t} \right), \tag{2.16}$$

con respecto a los coeficientes $\omega, \alpha_1, ..., \alpha_q$ y $\beta_1, ..., \beta_p$, donde ϕ denota la densidad normal estándar, y $\sigma_t = \sqrt{h_t}, t \geq 1$. Se hace recursivamente de (2.1.3) reemplazando ε_t por $\widetilde{\varepsilon}_t$, y con $\widetilde{\varepsilon}_t = 0$ y $h_t = \hat{\sigma}^2, t \geq 0$ para encontrar el mínimo de $-2\ln(L)$. Es importante que la optimización está restringida para que todos los parámetros estimados no sean negativos con

$$\hat{\alpha}_1 + ... + \hat{\alpha}_q + \hat{\beta}_1 + ... + \hat{\beta}_p < 1, \;\; \hat{\omega}_0 > 0 \tag{2.17}$$

La condición anterior es necesaria y suficiente para que la ecuación del correspondiente modelo $GARCH$ sea débilmente estacionario.

3. Caso de estudio: COVID-19 en Santander

3.1. Modelo ARIMA

Ahora antes de hacer el análisis con los datos reales, comprobaremos nuestra metodología con datos simulados. Primero vamos a generar los datos de una serie de tiempo basados en un proceso $ARIMA(5, 2, 3)$.

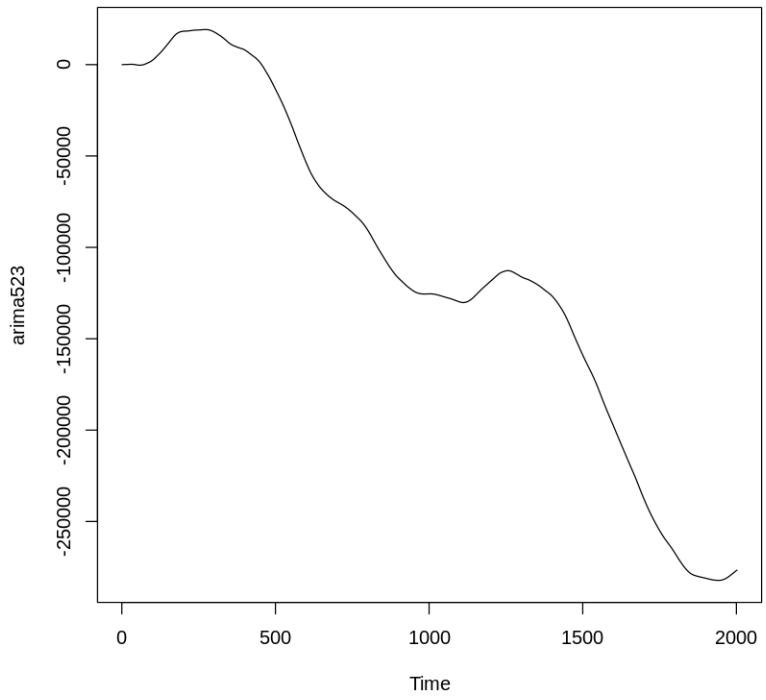


Figura 3.1: Datos simulados

Para encontrar la frecuencia adecuada usaremos un periodograma que se define de la siguiente manera.

Definición 3.1.1. [4] Se define el priodograma basado en muestras $X_{t_1}, X_{t_2}, ..., X_{t_n}$ como la función:

$$I(\lambda) = \frac{1}{\pi}\left(\hat{Cov}(X_{t_i}, X_{t_j}) + 2\sum_{h=1}^{n-1}\hat{Cov}(X_t, X_{t+h})cos(\lambda h)\right) \tag{3.1}$$

Con $\lambda \in [0, \pi]$ y para $t > 0$. El periodograma proporciona información sobre la contribución relativa de diferentes frecuencias a la serie de tiempo.

Enseguida queremos encontrar la frecuencia de los días en los cuales se puede observar un comportamiento estacional, entonces vamos a tomar las siguientes tres frecuencias: $s_1 = \frac{1}{0{,}0055} \approx 182$ en color azul, $s_2 = \frac{1}{72} \approx 72$ en color morado, y $s_3 = \frac{1}{0{,}002} = 500$ en color rojo.

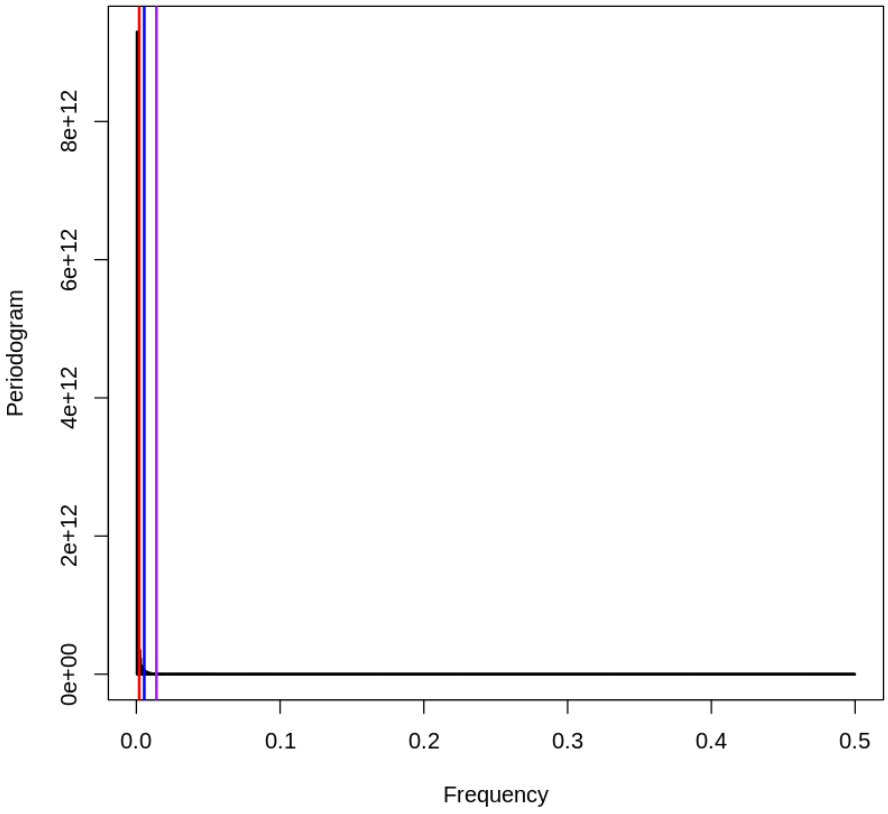


Figura 3.2: Periodograma datos simulados

Podemos que en la anterior gráfica la frecuencia s_2 esta muy alejada y no tiene prácticamente ningún pico, por otro lado la frecuencia s_3 aun teniendo un pico bastante alto es muy grande para ser considerada pensando en que nuestros datos son solamente 2000, por dicha razón tomaremos la frecuencia s_1. Suponiendo que en nuestros datos simulados las unidades del tiempo son días, vamos a analizar nuestros datos tomando intervalos cada 182 días como se muestra en la siguiente gráfica.

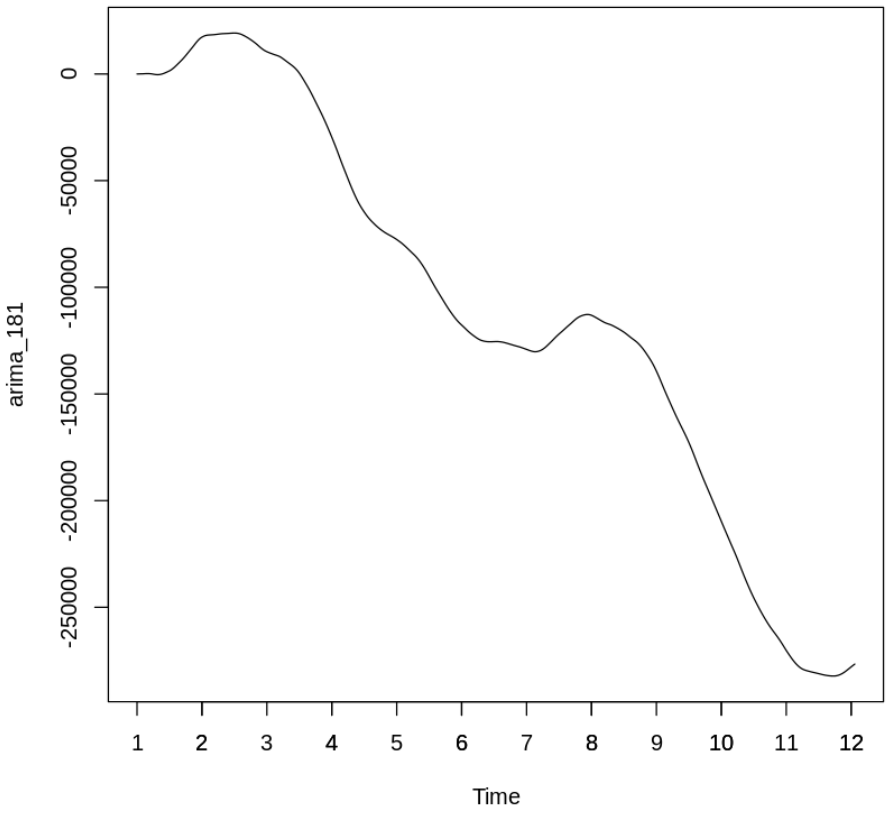


Figura 3.3: Datos periodo 182

Ahora bien, siguiendo con nuestra metodología debemos encontrar el orden de nuestro modelo $ARIMA(p, d, q)$, primero debemos saber que el numero de diferencias que debemos hacer son dos, es decir que $d = 2$, después para hallar los valores de p, q debemos usar las funciones de autocorrelación y autocorrelción parcial.

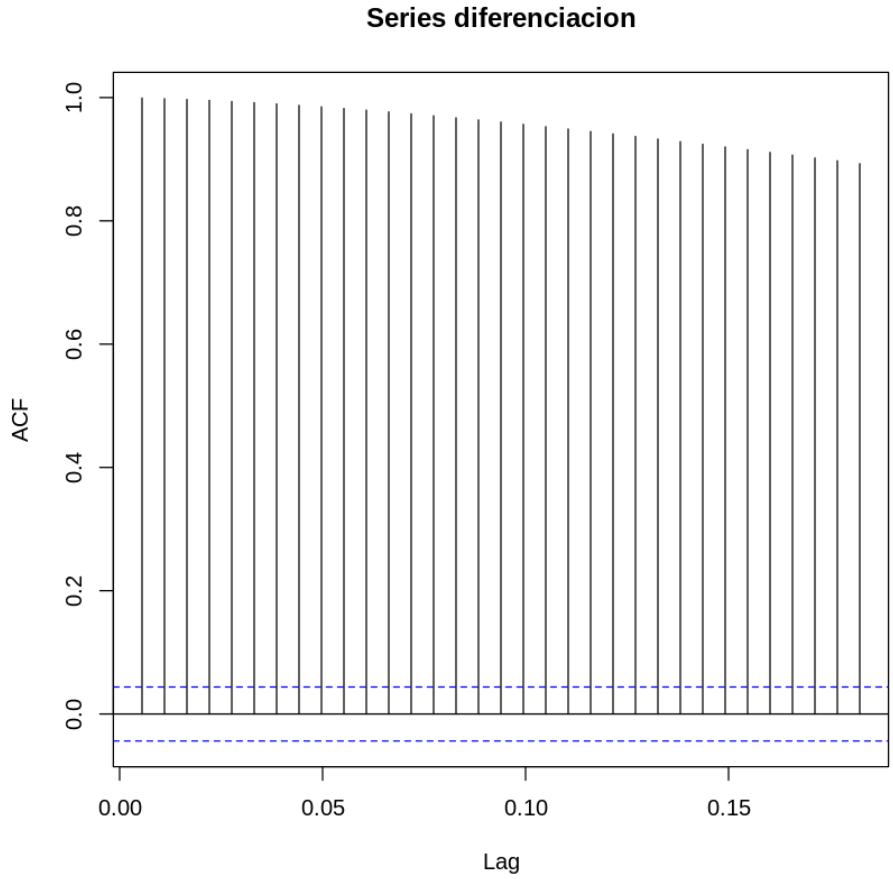


Figura 3.4: Auto correlación

Podemos ver que el numero de coeficientes es elevado entonces deberíamos todos el orden de q muy alto, sin embargo el seleccionar un orden muy alto no es recomendable ya que podríamos tener problemas de sobre ajuste o inestabilidad en nuestro modelo, usaremos el criterio de Akaike (1.27), para saber con mas precisión el orden de nuestro modelo. Para nuestro caso el orden $q = 5$.

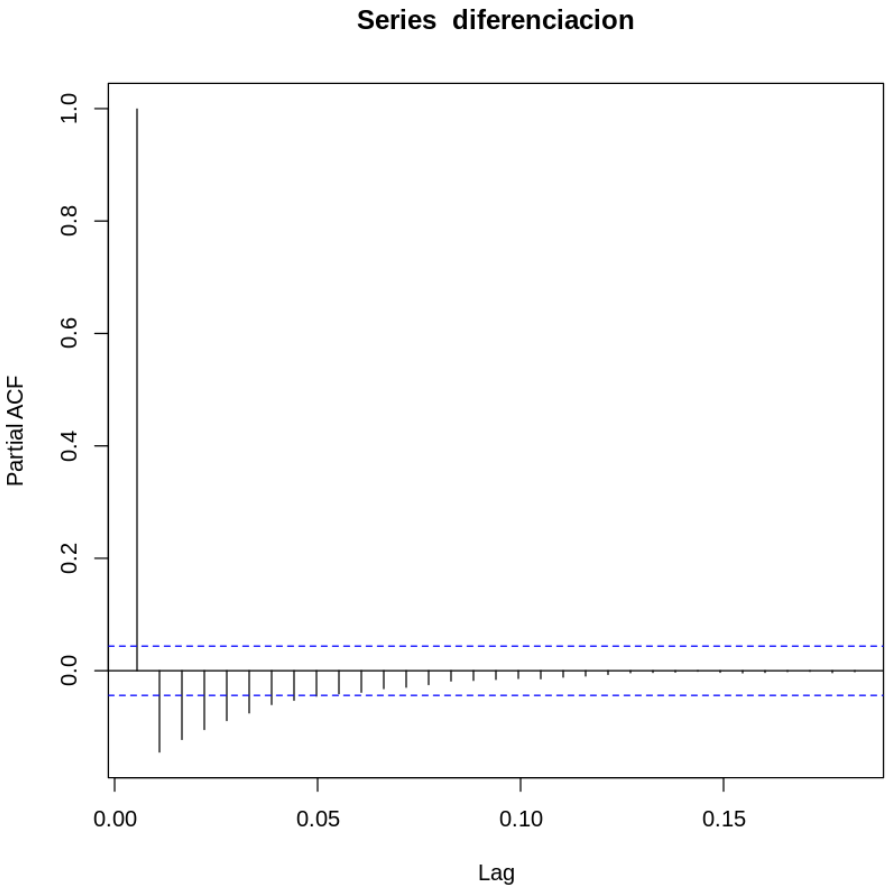


Figura 3.5: Auto correlación parcial

Para nuestro coeficiente p tenemos el mismo caso anteriores, tomaremos de la misma forma $p = 5$. Miremos ahora la gráfica para la predicción para los siguientes 100 valores de nuestro modelo. Podemos ver que la predicción para 100 futuros datos en la gráfica (3.6), como se puede ver la variación de los datos predichos no es muy alta con un comportamiento creciente. Ahora análisis los residuos de nuestro modelo que se muestran en la gráfica (3.7).

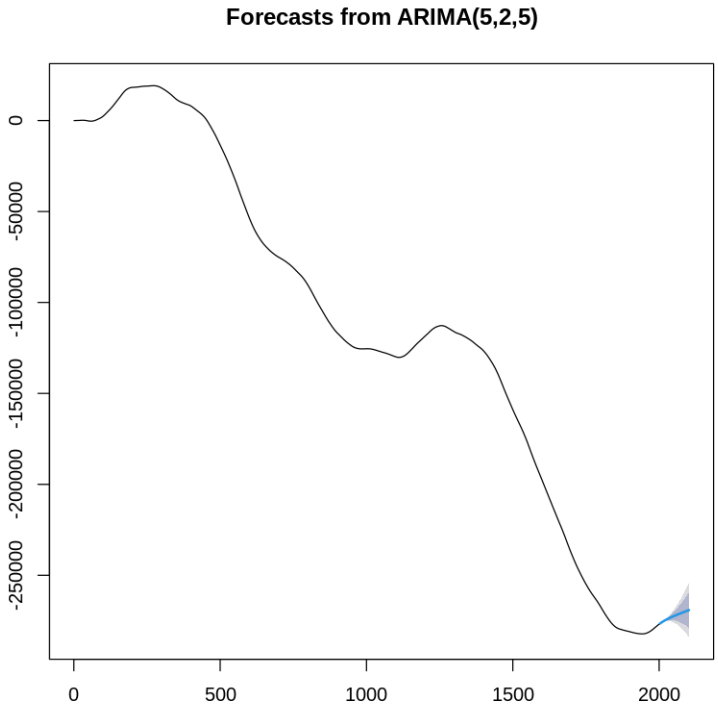


Figura 3.6: Predicción para 100 valores futuros

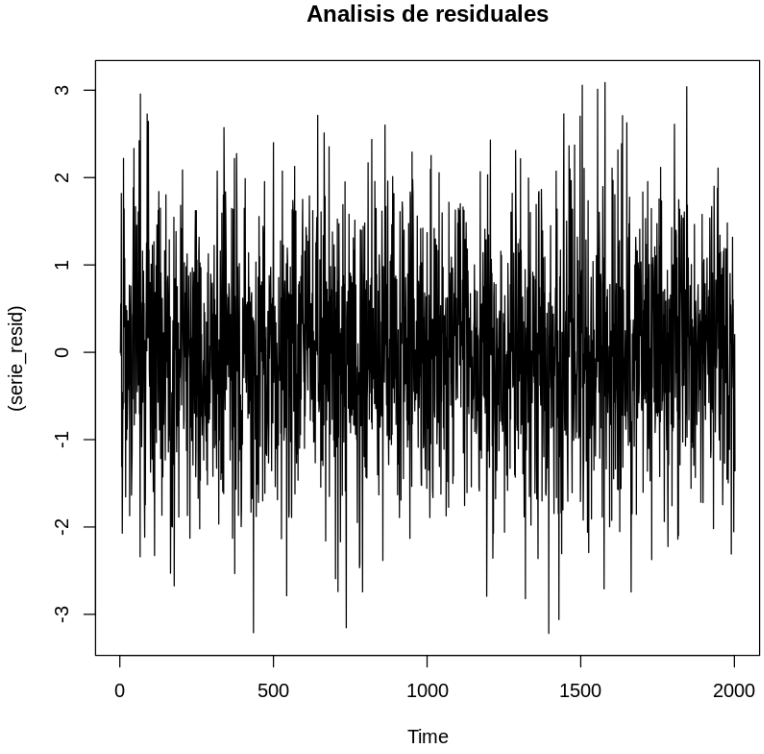


Figura 3.7: Residuos

Como podemos observar los errores de nuestro modelo no siguen algún tipo de patrón en especifico y tampoco se observa que se hacen mas pequeños a medida que pasa el tiempo, para poder confirmar que no existe un comportamiento heterocedastico usaremos la prueba de McLeodLi con un nivel de significancia de $\alpha = 0{,}05$.

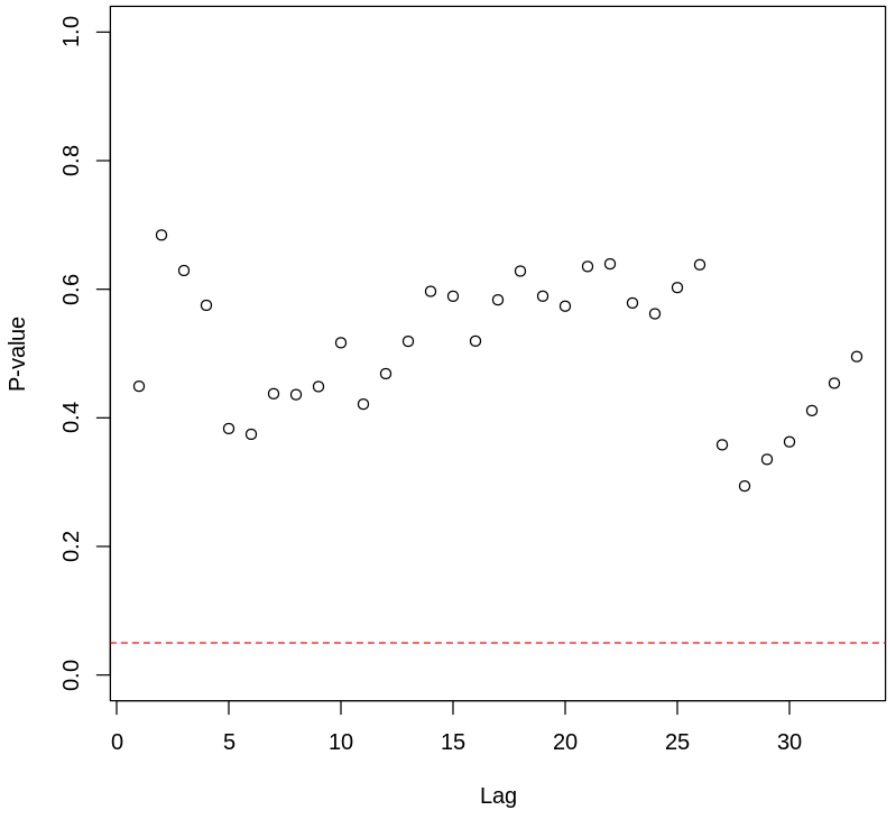


Figura 3.8: Prueba de McLeodLi

se puede ver que para cada retardo su autocorrelación parcial de los errores están por encima del nivel de confianza de α, entonces no podemos rechazar la hipótesis nula de la no heterocedasticidad por tanto hay homocedasticidad, y tiene sentido con nuestra metodología ya que los datos son generados por un proceso $ARIMA$, que tiene se tiene el supuesto de homocedasticidad.

De este ejemplo podemos concluir que bajo la siguiente metodología, se obtienen

órdenes adecuados para la parte autorregresiva y de medias móviles.

- 1. Se realiza un periodograma con el fin de encontrar el periodo, es decir, en qué momentos se empieza a repetir el comportamiento de la serie.

- 2. Se diferencia la seria para determinar el orden de integración estacionario, d.

- 3. Se escogen los órdenes pyq, los cuáles son los órdenes estacionarios. Con la función de autocorrelación parcial se escoge p a partir del punto donde esta tiende a cero. Con la función de autocorrelación se escoge q a partir del punto donde esta tiende a cero.

- 4. Para los órdenes P, Q, D, los cuáles son estacionales, se utiliza el criterio de Akaike y se escoge el modelo con el mejor ajuste.

El codigo utilizado para dicha simulción fue el siguiente.

`https://github.com/CrasCris/Ejemplo1/blob/main/Simulacin.ipynb`

3.2. Descripción de la base de datos

De acuerdo con [19], existen diversidad de pruebas y metodologías en el laboratorio que en conjunto con la definición de caso, sintomatología y antecedentes epidemiológicos sirven en la detección y diagnóstico de la enfermedad, causada por el nuevo Coronavirus SARS-CoV-2 del 2019 (COVID-19). Las pruebas de laboratorio pueden estar basadas en la detección directa de la presencia del virus o en la detección indirecta mediante métodos serológicos detectando producción de anticuerpos como respuesta a la infección. Para detectar la presencia del virus se usa principalmente la RT-PCR en tiempo real que identifica ARN viral específico de SARS-CoV-2 y para la detección de anticuerpos se

[19] Instituto Nacional de Salud. *Pruebas para la detección molecular de SARS-COV-2 por RT-PCR usadas en Colombia*. Url: `http://www.ins.gov.co/BibliotecaDigital/Pruebas-deteccion-molecular-sars-cov-2-rt-pcr-Colombia.pdf`. 2021.

registra el uso de tres metodologías: i) quimioluminiscencia (CLIA), ii) inmunoabsorción ligado a enzimas (ELISA) y iii) inmunocromatografía (pruebas rápidas en casete), entre otras. Sin embargo, aún no es totalmente claro el debido uso de estas últimas, por consiguiente, el presente documento pretende resumir parte de la información disponible a la fecha que sirva como evidencia para un mejor uso de estas[20].

Se realizaron reuniones con el director para aclarar dudas acerca de los resultados obtenidos. Se realizara primero el análisis con un modelo ARIMA, para este haremos una simulación para confirmar nuestra metodología, y posteriormente con un modelo GARCH. Se utilizará los paquetes `stats`, `forecast` y `rugarch` para la implementación de los modelos sobre los datos, cuya descripción se aborda a continuación:

La base de datos de COVID-19 en Colombia es tomada de la pagina del Instituto Nacional de Salud [21]. La base de datos es de uso publico y se actualiza de manera semanal cada jueves. En donde se encuentra la especificación y aclaración formal de las siguientes variables:

- `Fecha de reporte vía web`, es la fecha indicada en [21].
- `ID de caso`, un código único para caso de contagio .
- `Fecha de notificación`, fecha del día que se notifica el contagio.
- `Codigo DIVIPOLA`, es una nomenclatura estandarizada, diseñada por el DANE para la identificación de Entidades Territoriales (departamentos, distritos y municipios), Áreas No Municipalizadas y Centros Poblados, mediante la asignación de un código numérico único a cada una de estas unidades territoriales.[22]
- `Nombre del departamento`.
- `Código DIVIPOLA del municipio`.

---

[20] https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx

[21] Instituto Nacional de Salud. *Base de datos*. Url: `https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data`. 2023.

[22] Departamento Administrativo Nacional de Estadistica. *Significado del codigo DIVIPOLA*. Url: `https://www.ins.gov.co/Noticias/paginas/coronavirus.aspx`. 2023.

- `Edad`, la edad que tenia la persona a la fecha cuando se tomo la prueba .
- `Unidad de medida de edad`, que corresponde a 1-años, 2-meses, 3-dias.
- `Sexo` Masculino(M) o femenino(F).
- `Tipo de contagio`, variable que puede tomar los valores de Çomunitario",Relacionado",etc.
- `Ubicación del caso`, toma los valores de "Hospital UCI", "Fallecido", o "N/A", corresponde a muertes no relacionadas con COVID-19, aún si eran casos activos.Hay pacientes recuperados para COVID-19, que pueden permanecer en hospitalización por otras comorbilidades.[21]
- `Estado`, variable que puede tener los valores de "Leve", "Moderado", "Grave", "Fallecido", o "N/A". Es cuando no se tiene registro del estado de la persona.
- `Codigo ISO del país`, se define como un estándar internacional para los códigos de país y los códigos para sus subdivisiones. Ha sido publicada por la Organización Internacional de Normalización [23]. Solamente aplica para los casos de tipo Ïmportado".
- `Nombre del país` de donde fue importado el contagio.
- `Recuperado`, Recuperado Fallecido N/A (Vacío). N/A se refiere a los fallecidos no COVID. Pueden haber casos recuperados con ubicación Hospital u Hospital UCI, ya que permanecen en hospitalización por causas diferentes. Los casos con información en blanco en esta columna corresponde a los casos activos.[21]
- `Fecha de inicio de síntomas` , es la fecha donde se estima la persona inicio con los sintomas como (fiebre,malestar general,tos,etc).
- `Fecha de muerte`, es la fecha declara en la que el individuo fallece , puede variar respecto a la real.
- `Fecha de diagnostico`, es la fecha que se reporta cuando la prueba da positivo.
- `Fecha de recuperación`, fecha cuando se reporta que la persona se ha recuperado.

---

[23] Organizacion Nacional de Organizacion. *Significado del codigo ISO*. 2023.

- `Tipo de recuperación`, Se refiere a la variable de tipo de recuperación que tiene dos opciones: PCR y tiempo. PCR indica que la persona se encuentra recuperada por segunda muestra, en donde dio negativo para el virus; mientras que tiempo significa que son personas que cumplieron 30 días posteriores al inicio de síntomas o toma de muestras que no tienen síntomas, que no tengan más de 70 años ni que estén hospitalizados.[21]

- `Pertenencia étnica` ,1-Indígena 2-ROM 3-Raizal 4-Palenquero 5-Negro 6-Otro. Esta variable se actualizará cada semana.[21]

- `Nombre del grupo étnico`.

Los datos a modelar corresponden al conteo diario de `Fecha de diagnostico` (como número diario de infectados), ya que para la variable de `Fecha de inicio de síntomas` no tiene en cuenta los pacientes asintomáticos, y, por otra parte, no todos los diagnosticados se conoce la fecha exacta de inicio de síntomas.

3.2.1. Analisis COVID-19 en Satander [7]Ahora vamos a implementar el modelo ARIMA usando la metodología que previamente mencionamos con los datos de COVID-19 en el departamento de Santander obtenidos de la pagina del Ministerio de salud y protección social [21] desde marzo 3 del 2020 al 19 de abril del 2023, el número reproductivo básico es mayor a 1 por lo que se deben tomar todos los datos para determinar el periodo de tiempo bajo el cual se tienen picos para la enfermedad [24], nuestra variable a considerar es la fecha de reporte de inicio de síntomas.

---

[24] David Niño-Torres et al. “Stochastic modeling, analysis, and simulation of the COVID-19 pandemic with explicit behavioral changes in Bogotá: A case study”. En: *Infectious Disease Modelling* 7.1 (2022), págs. 199-211.

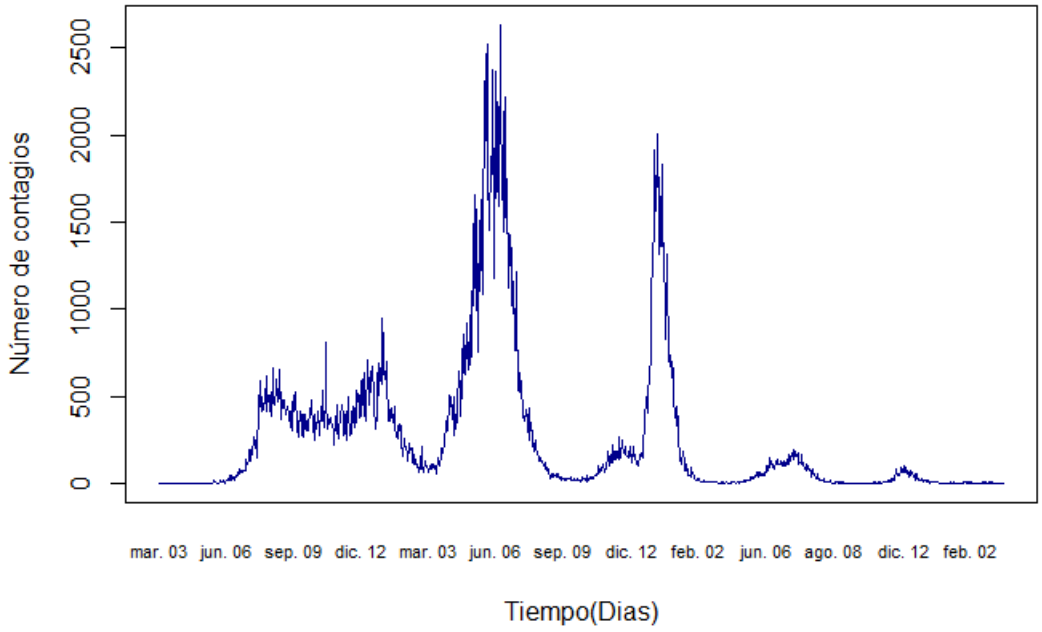


Figura 3.9: Gráfica de los contagios

Dado que nuestros datos no son regulares, es decir, existen días en los cuales no hubo reporte de la cantidad de personas contagiadas, por tanto debemos recurrir a una interpolación spline para tener una serie de tiempo regular, el tomar este tipo de interpolación particular no es un capricho, ya que el algoritmo spline para interpolar datos usa polinomios de grado inferior para aproximar los datos lo que nos garantiza una mínima perturbación de la forma general de la serie de tiempo garantizando el mínimo error acumulado. Usando las librerías:

- **[**TSA][10]: Las funciones `periodogram()` para encontrar la frecuencia adecuada, la función `McLeod.Li()` para realizar la prueba de McLeod Li sobre la no heterocedasticidad en los errores.

- **[**Forecast][25]: La función `forecast()` para realizar las predicciones.

[25] Rob Hyndman y et al. *forecast: Forecasting functions for time series and linear models*. R package version 8.21. 2023. URL: `https://pkg.robjhyndman.com/forecast/`.

- **[**Stats]**[26]**: Las funciones `ts()` para crear objetos del tipo serie de tiempo, `acf()` la gráfica de la función de autocorrelación, `pacf()` la gráfica de la función de autocorrelación parcial, `AIC()` calcular el criterio de Akaike.

- **[**ImputeTS]**[27]**: La función `na.interpolation()` para hacer la interpolación de los datos faltantes usando el método de spline.

Ahora vamos a usar un periodograma para poder encontrar la frecuencia en la cual podemos observar un comportamiento estacional en nuestra serie de tiempo usando la metodología explicada en [28], dependiendo de los picos que presente nuestro periodograma y en que frecuencia se encuentra dicho pico, en la gráfica se escogieron la frecuencia de $s_1 = \dfrac{1}{0{,}0055} \approx 181$ en color azul y $s_2 = \dfrac{1}{0{,}014} \approx 71$ en color morado.

---

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/.

[27] Steffen Moritz y Thomas Bartz-Beielstein. "imputeTS: Time Series Missing Value Imputation in R". En: *The R Journal* 9.1 (2017), págs. 207-218. DOI: 10.32614/RJ-2017-009. URL: https://doi.org/10.32614/RJ-2017-009.

[28] Santiago De la Fuente. *Series temporales, modelo ARIMA, metodología de Box-Jenkins*. Facultad de Ciencias Económicas y Empresariales, Departamento de Economía Aplicada, Universidad Autónoma de Madrid. Enlace web: https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf. 2022.

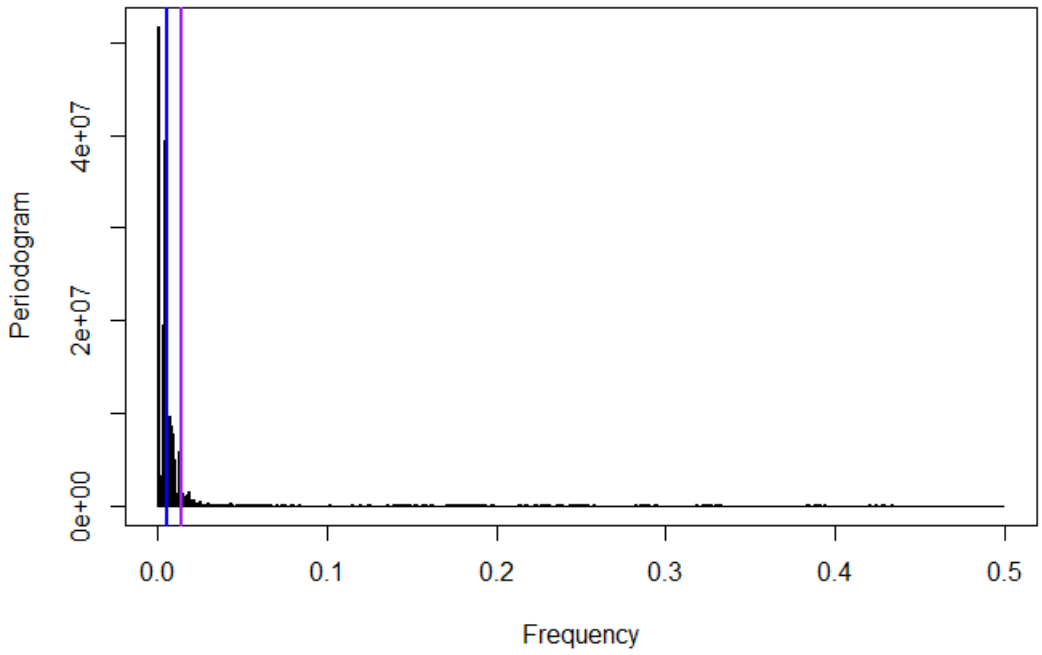


Figura 3.10: Periodograma

Podemos ver en la gráfica del periodograma en los primeros periodos hay picos, sin embargo su frecuencia es muy baja, por otra parte podemos ver que en la frecuencia de $s_2 \approx 71$ tiene un pico relativamente pronunciado, por tanto vamos a considerar esa nuestra frecuencia. Ahora teniendo nuestra frecuencia establecida podemos seguir con su análisis.

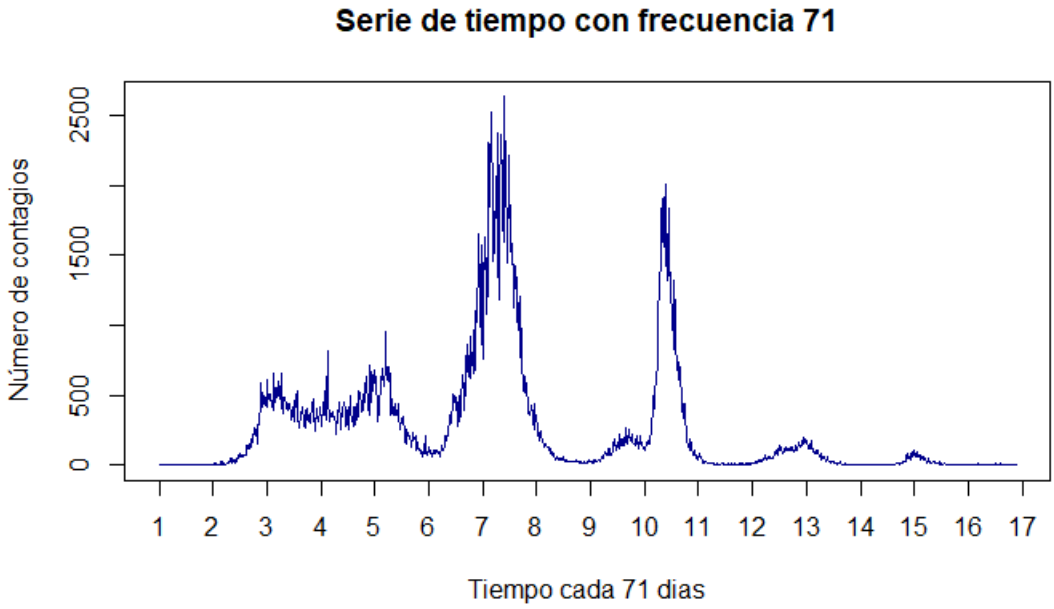


Figura 3.11: Gráfica con periodo 71

Como podemos ver en la gráfica 3.11 con periodo en 71 se realizo con el siguiente código:

```
serie_inicio_2<-ts(serie_inicio,frequency = 71)
#Siendo serie_inicio nuestro datos de número de contagios
#GRAFICA DE LA SERIE DE TIEMPO CON FRECUENCIA DE 71
plot(serie_inicio_2,xaxt="n",main="Serie de tiempo con frecuencia 71",
col="darkblue",xlab="Tiempo cada 71 dias",ylab="Número de contagios")
axis(side = 1,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17))
```

consideramos los momentos mas importantes de serie de tiempo, los cuales serian sus picos mas alto y los momentos donde hay un cambio muy grande de su varianza, ahora siguiendo con nuestra metodología queremos saber cuantas diferencias hacen falta para lograr un comportamiento estacional, que en nuestro caso tiene un valor de $d = 0$. Debemos también encontrar los coeficientes q y p respectivamente usando las funciones de autocorrelación y autocorrelación parcial viendo a partir de que valor se van haciendo cero, que dependiendo la cantidad de puntos fuera de la linea punteada tomada como valores cercanos a cero, dado que según nuestra metodología estamos considerando la parsimonia es decir entre menos los ordenes de nuestro modelo.

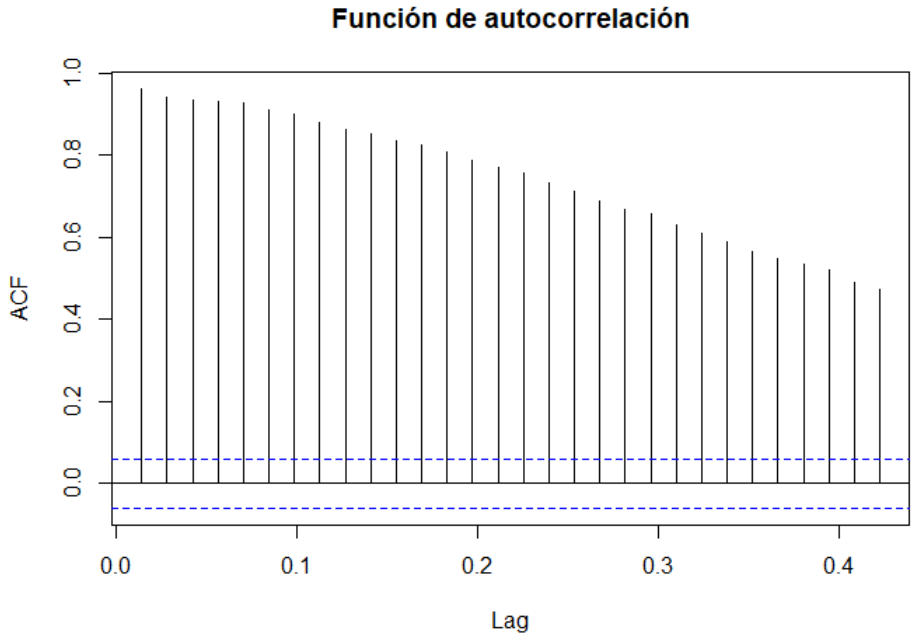


Figura 3.12: Gráfica función de autocorrelación

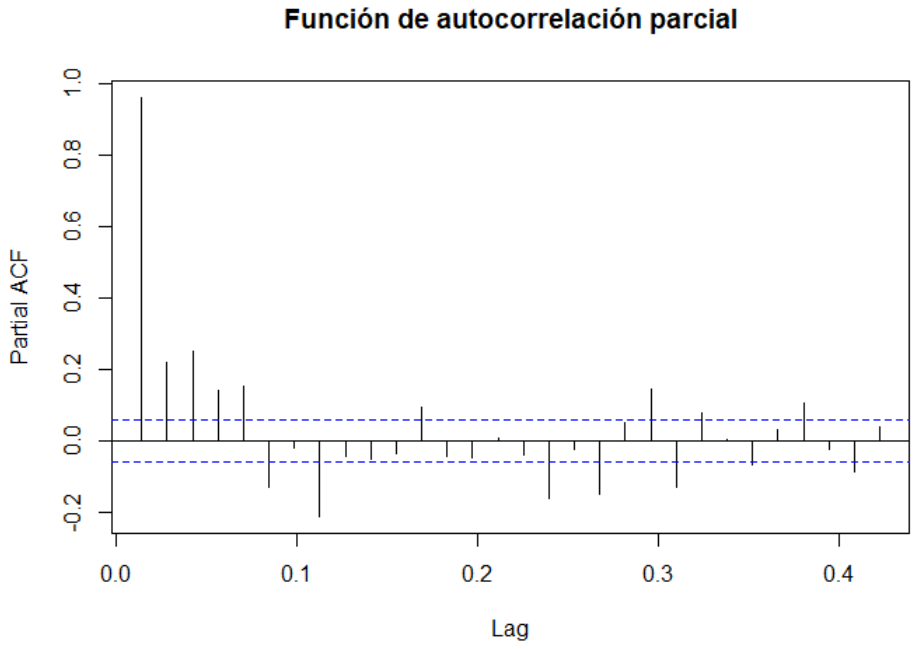


Figura 3.13: Gráfica función de autocorrelación parcial

Dichas gráficas y el calculo de la diferencia d se realizaron con el siguiente código:

```
nsdiffs(serie_inicio_p2)#
pacf(serie_inicio_p2,main="Función de autocorrelación parcial")
acf(serie_inicio_p2,main="Función de autocorrelación")
```

Como podemos ver en la primera gráfica (3.12) tendríamos que nuestro orden para la parte de media móviles sería muy elevado, siguiendo con lo mencionado antes de realizar las gráficas vamos a tomar este orden como $q = 5$, por otra parte para la segunda gráfica (3.13) podemos ver que también tendríamos un orden elevado, sin embargo los valores con mayor relevancia son 5, por tanto tomaremos el orden para la parte de autorregresiva $p = 5$. Además el número de diferencias es decir su orden de la parte integral para nuestra parte estacionaria es de $d = 0$.

Ahora ya tenemos nuestros coeficientes para la parte regular de nuestro modelo, miremos ahora los coeficientes para la parte estacional, para encontrar dichos coeficientes usaremos el criterio de Akaike (1.27), vamos a encontrar los coeficientes P, D, Q de nuestro modelo $ARIMA(5, 0, 5) \times (P, D, Q)$, dado que el costo computacional de crear cada modelo y encontrar su correspondiente criterio de Akaike, evaluaremos $P, Q = 0, 1, 2, 3$ y $D = 0, 1, 2$.

Tenemos entonces los siguientes tres modelos con mejor criterio de Akaike para cada iteración.

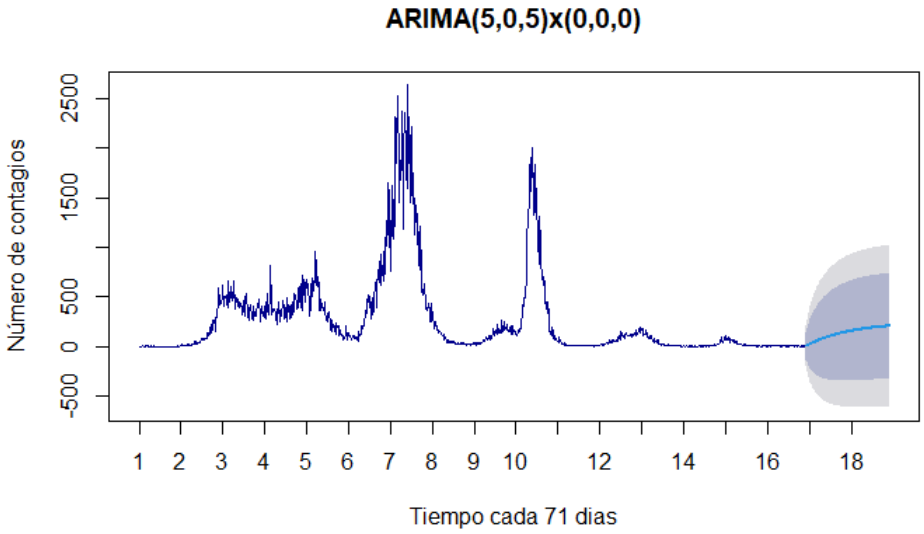


Figura 3.14: Gráfica del tercer modelo con mejor criterio

Podemos ver que de este primero modelo (3.14) el cual es que el tiene menor criterio de Akaike para ser especifico de $AIC = 13203{,}59$, podemos ver que en nuestra predicción con este modelo tiene un comportamiento creciente. También resaltar que hacemos la proyección para 2 ciclos de 71 días, es decir aproximadamente 5 meses. De la misma forma dado que la banda de confianza sugiere una posibilidad de número de contagios negativo, pero nos enfocaremos en la parte del número de contagios positivos, veamos el siguiente modelo.

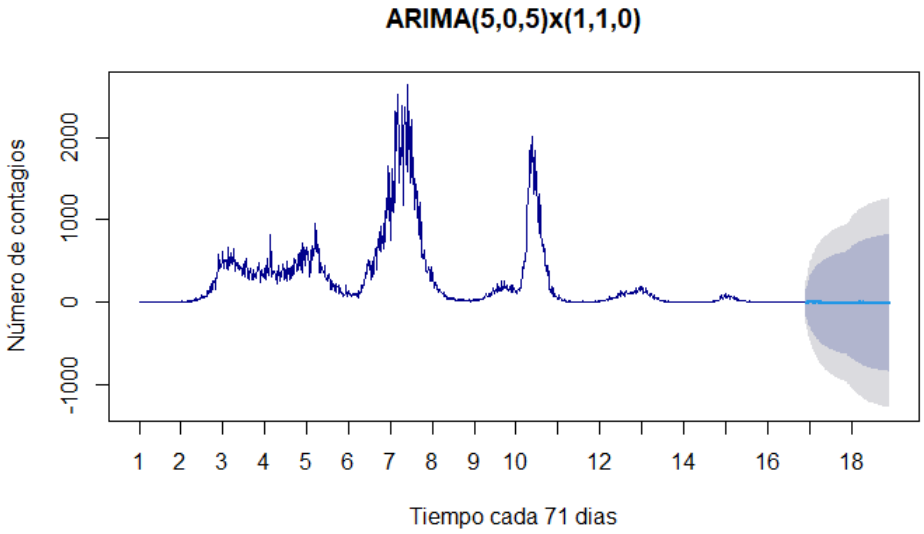


Figura 3.15: Gráfica del segundo modelo con mejor criterio

Como podemos ver en la gráfica (3.15) ya no presenta un crecimiento tan pronunciado como el modelo anterior, por otra parte podemos apreciar que el componente para la parte estacional de la media móvil $MA(q)$ sigue siendo $q = 0$ al igual que el modelo anterior, en este modelo ya se puede apreciar con comportamiento estacional , cuenta con un valor en su criterio de Akaike de $AIC = 12527{,}77$, podríamos considerar este modelo lo suficientemente acertado considerando como antecedente el modelo anterior, tiene una mejor de $675{,}82$ en su criterio de Akaike, veamos ahora que tanto mejora el siguiente modelo frente al segundo.

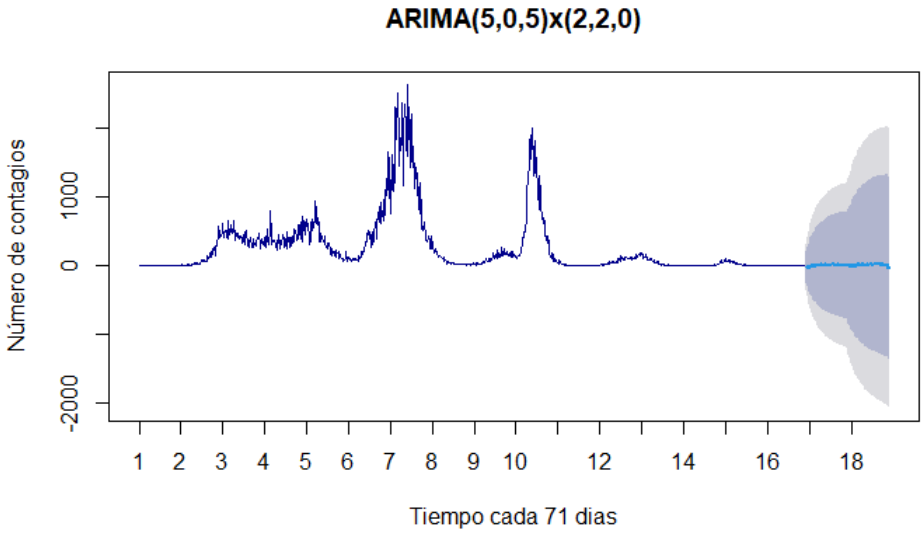


Figura 3.16: Gráfica del mejor modelo con mejor criterio

Podemos observar en la gráfica 3.16 que a simple vista no se nota una diferencia notoria frente al modelo anterior y también lo podemos confirmar comparando su criterio de Akaike, que para este modelo es de $AIC = 12111{,}31$, y tiene una mejor frente al modelo anterior de solo $416{,}46$, tenemos entonces que nuestro mejor modelo es,

$$ARIMA(5,0,5) \times (2,2,0)_{71}. \tag{3.2}$$

Analicemos ahora el comportamiento de los errores de nuestro modelo, para ello vamos a usar el test de [29] del paquete **[**TSA]10, que nos dice si existe un comportamiento condicional heterocedastico o no en nuestro modelo, para una confianza del $95\,\%$ tenemos la siguiente gráfica, también vamos a analizar la gráfica de los residuales en el tiempo

[29] Allan I McLeod y William K Li. "Diagnostic checking ARMA time series models using squared-residual autocorrelations". En: *Journal of time series analysis* 4.4 (1983), págs. 269-273.

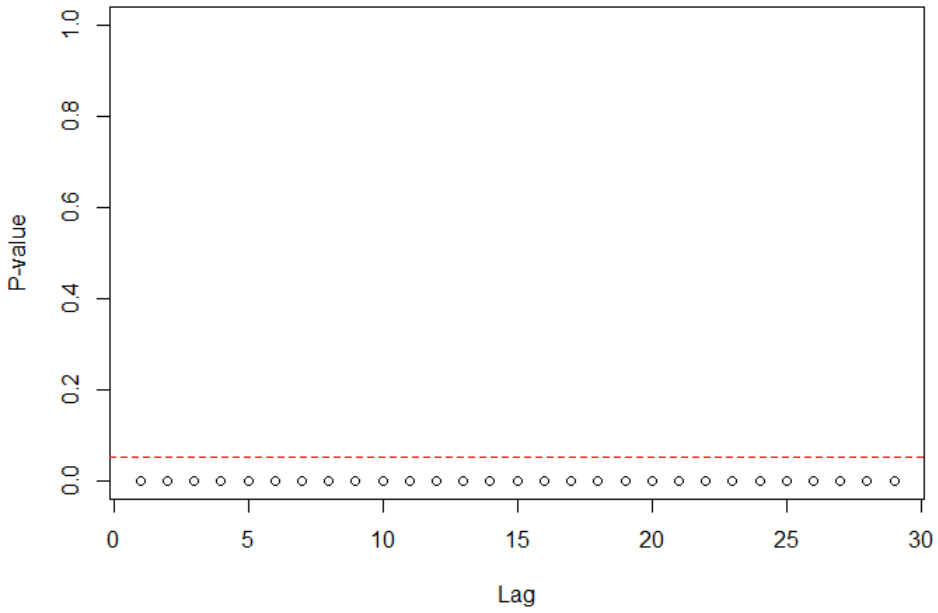


Figura 3.17: Gráfica McLeod-Li

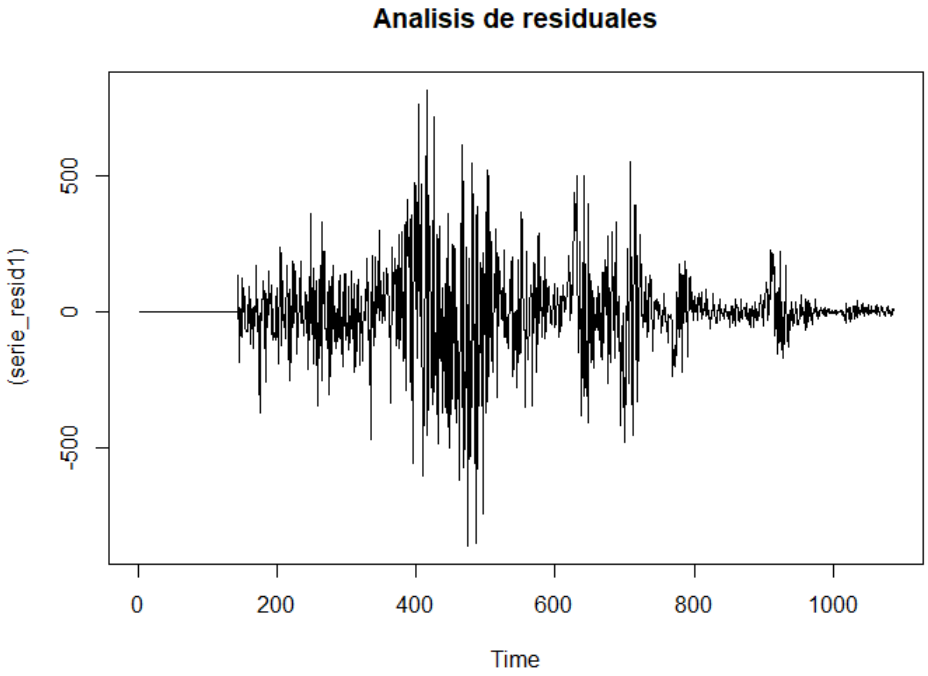


Figura 3.18: Gráfica Residuales

Podemos ver que en la gráfica (3.17) los retardos quedan por debajo de un nivel de confianza del $\alpha = 0{,}05$, por tanto rechazamos la hipótesis de heterocedasticidad, es decir, la varianza de los errores es constante a lo largo del tiempo, miremos ahora la gráfica de los residuales con una confianza del $95\,\%$

A pesar de que inferencialmente a través del test de Mc-Leod Li se encuentra homocedasticidad, de acuerdo con la grafic (3.18) si hay heterocedasticidad.

3.3. Modelo GARCH

Ahora para nuestro análisis usaremos la siguiente **[**rugarch]**[30]** con las funciones `ugarchspec()` que crea un objeto de tipo serie de tiempo $GARCH$, `ugarchfit()` que es para ajustar la varianza de un modelo , `ugarchforecast()` para la predicción.

---

[30] Tobias Kley Alexios Galanos. *Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. 2022. URL: `https://cran.r-project.org/web/packages/rugarch/index.html`.

$Orden(p,q)$	$sGARCH$ 2	$eGRCH$
$(0,0)$	$-6390{,}251$	$-6390{,}251$
$(1,0)$	$-6486{,}153$	$\times$
$(1,1)$	$-5100{,}143$	$-5046{,}584$
$(0,1)$	$-6389{,}533$	$-6390{,}254$
$(2,0)$	$-5373{,}551$	$\times$
$(2,1)$	$-5100{,}995$	$-5049{,}691$
$(2,2)$	$-5102{,}868$	$-5049{,}39$
$(0,2)$	$-6390{,}175$	$-6390{,}251$
$(1,2)$	$-5100{,}139$	$-5071{,}986$
$(3,0)$	$-5315{,}018$	$\times$
$(3,1)$	$-5080{,}519$	$-5040{,}341$
$(3,2)$	$-5085{,}493$	$-5040{,}252$
$(3,3)$	$-5085{,}494$	$-5049{,}839$
$(0,3)$	$-6390{,}244$	$-6390{,}251$
$(1,3)$	$-5070{,}932$	$-5038{,}222$
$(2,3)$	$-5097{,}443$	$-5129{,}918$
$(4,0)$	$-5140{,}045$	$\times$
$(4,1)$	$-5065{,}214$	$-5038{,}851$
$(4,2)$	$-5065{,}217$	$-5136{,}077$
$(4,3)$	$-5065{,}207$	$-5034{,}243$
$(4,4)$	$-5065{,}205$	$-5046{,}752$
$(0,4)$	$-6368{,}752$	$\times$
$(1,4)$	$-5100{,}139$	$-5049{,}417$
$(2,4)$	$-5097{,}443$	$-5027{,}317$
$(3,4)$	$-5080{,}518$	$-5036{,}351$
$(5,0)$	$-5094{,}101$	$\times$
$(5,1)$	$-5065{,}215$	$-5033{,}922$
$(5,2)$	$-5069{,}06$	$-5050{,}046$
$(5,3)$	$-5064{,}807$	$-5023{,}303$
$(5,4)$	$-5064{,}806$	$-5046{,}069$
$(5,5)$	$-5064{,}808$	$-5051{,}955$
$(0,5)$	$-6389{,}809$	$\times$

Tabla 3.1: Criterios de Akaike

Ahora ya tenemos los coeficientes para la parte estacionaria que encontramos de la parte $ARIMA$ (3.2) que recordando los coeficientes fueron $p = 5$ y $q = 5$, para encontrar los coeficientes de nuestro modelo $GARCH$, miraremos que modelo maximiza la función de máxima verosimilitud, realizaremos la iteración del modelo para 2 tres tipos distintos de modelo; $sGARCH$ o el modelo $GARCH$ estándar y $eGARCH$ o el modelo $GARCH$ exponencial, en donde vamos a variar los ordenes correspondientes $p, q = 0, ..., 5$ y analizaremos cual de cada uno de ellos consigue maximizar la función de máxima verosimilitud.

En nuestro caso tenemos que en la tabla (3.1) existen algunos problemas de convergencia, podemos ver que el modelo $sGARCH(1, 0)$ es el que más maximiza la función de máxima verosimilitud, para nuestro caso tenemos que la gráfica de la predicción para el Sigma es la siguiente,

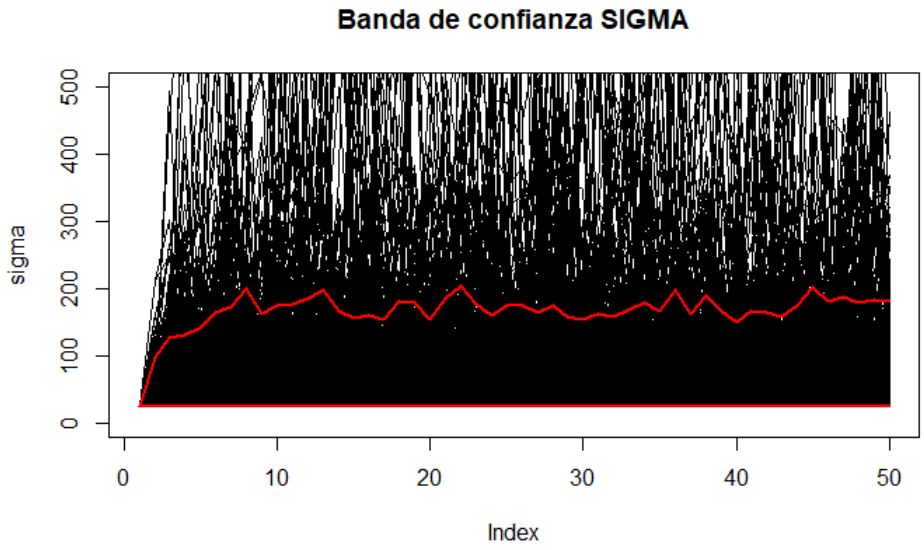


Figura 3.19: Gráfica banda Sigma

Podemos ver en la gráfica (3.19) para una predicción de 50 días el comportamiento del Sigma, se pueden ver de color rojo el primer y ultimo cuantil respectivamente, observar que el comportamiento del sigma para cada uno de los días no es constante, cada uno de los días varia, esto nos dice que la volatilidad para cada día es distinta y varia para cada día. Miremos ahora la predicción para el valor de la serie en esos 50 días.

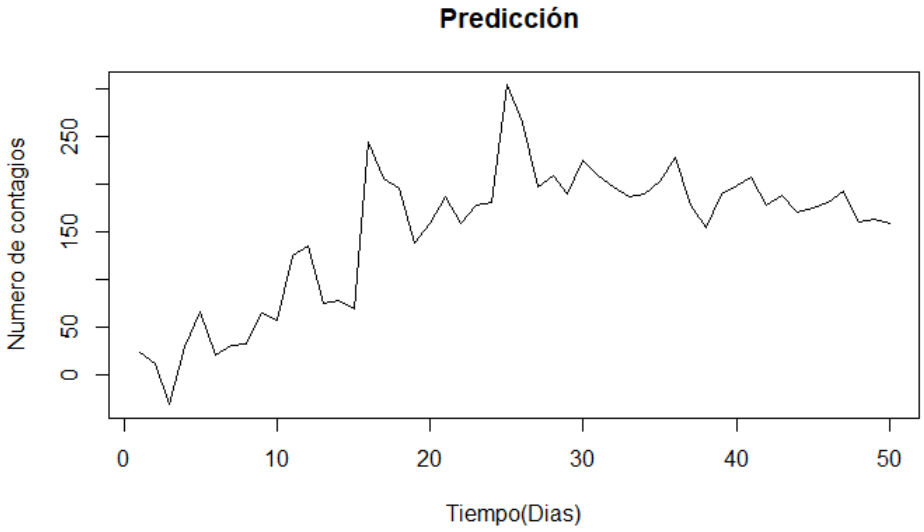


Figura 3.20: Gráfica predicción

Se puede que los valores de la serie tiene picos bastantes pronunciados donde su varianza es distinta para cada momento, contrastando con la predicción para los valores de nuestro modelo (3.2), no obstante podemos ver que al final de la predicción se puede apreciar que los valores van reduciendo poco a poco.

4. Conclusiones

Según los objetivos establecidos en este estudio y después de examinar los hallazgos obtenidos, se exponen las siguientes conclusiones:

En nuestro análisis pudimos ver la predicción del número de casos de covid-19 en Santander a través del modelo $ARIMA$ en la figura (3.16). Se encontró que el número de casos positivos no aumento significativamente, por lo cual, no representa el virus un problema para la salud publica. Respecto al modelo $GARCH$ observado en la figura (3.19), se predice que la variabilidad de los casos de un día respecto a otro tendrá ciertos picos, pero no se considera que la variabilidad del número de casos sea significativa de un día a otro.

Por otra parte existen limitaciones técnicas respecto a una modelización con series de tiempo, sobre todo en la fase de estimación, así mismo otras variantes de los modelos $ARIMA$ y $GARCH$, como por ejemplo $ARFIMA$ Modelo Autorregresivo de Media Móvil Fraccionalmente Integrado usando la misma libreria `Forecast`, $IGARCH$ Modelo GARCH Integrado que se encontra en la librería `rugarch`, de igual forma contrastar las predicciones con modelos de ecuaciones diferenciales o usando inteligencia artificial.

Las librerías usadas fueron: `TSA` que fue la utilizada de la creación de las series de tiempo ARIMA, su limitación radica en la compatibilidad de los objetos que crea al combinarla con otras librerías, por ejemplo la librería `Forecast`, donde se debe tener especial cuidado a la de generar la predicción, y la librería `stats` que contiene estadísticos de prueba ; `rugarch` utilizada para la creación de los modelos GARCH, su principal limitación fue el calculó de los estadísticos de prueba que en muchas ocasiones presentaba multiples errores de convergencia al calcular el criterio de Akaike.

La metodología manejada corresponde a primero realizar el periodograma para encontrar el periodo adecuado, despues identificar el orden de integración d y los ordenes p, q de la parte estacionarios, y por ultimo usando el criterio de Akaike los ordenes de la parte estacionales P, Q, D para el modelo de mejor ajuste, la cual es adecuada al usar datos simulados.

Código

https://github.com/CrasCris/Proyecto/blob/master/ProyectoFinal.R

Bibliografía

Alexios Galanos, Tobias Kley. *Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. 2022. URL: https://cran.r-project.org/web/packages/rugarch/index.html (vid. pág. 69).

Bollerslev, Tim, Robert F Engle y Daniel B Nelson. "ARCH models". En: *Handbook of econometrics* 4 (1994), págs. 2959-3038 (vid. págs. 40, 41).

Box, George EP y David A Pierce. "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models". En: *Journal of the American statistical Association* 65.332 (1970), págs. 1509-1526 (vid. pág. 33).

Brockwell, Peter J y Richard A Davis. *Time series: theory and methods*. Springer science & business media, 2009 (vid. págs. 20, 22, 46, 58).

Cardona, Ortiz. "Propuesta de modelo ARIMA para la serie temporal de los casos de Covid-19 en Colombia aplicando la metodología Box and Jenkins". En: (2020) (vid. pág. 12).

Castañeda, Liliana Blanco, Viswanathan Arunachalam y Selvamuthu Dharmaraja. *Introduction to probability and stochastic processes with applications*. John Wiley & Sons, 2012 (vid. pág. 15).

Cavanaugh, Joseph E. "Unifying the derivations for the Akaike and corrected Akaike". En: (1996) (vid. pág. 31).

Chan, Kung-Sik y Brian Ripley. *TSA: Time Series Analysis*. 2022. URL: https://cran.r-project.org/package=TSA (vid. págs. 32, 59, 67).

Estadistica, Departamento Administrativo Nacional de. *Significado del codigo DIVIPOLA*. Url: https://www.ins.gov.co/Noticias/paginas/coronavirus.aspx. 2023 (vid. pág. 56).

Fisher, Ronald A. “On the mathematical foundations of theoretical statistics”. En: *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.594-604 (1922), págs. 309-368 (vid. pág. 30).

Fuente, Santiago De la. *Series temporales, modelo ARIMA, metodología de Box-Jenkins*. Facultad de Ciencias Económicas y Empresariales, Departamento de Economía Aplicada, Universidad Autónoma de Madrid. Enlace web: `https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf`. 2022 (vid. pág. 60).

Ghalanos, Alexios. “Introduction to the rugarch package.(Version 1.3-1)”. En: *Manuscript, http://cran. r-project. org/web/packages/rugarch. Accessed* 11 (2020) (vid. pág. 42).

González Casimiro, María Pilar. “Análisis de series temporales: Modelos ARIMA”. En: (2009) (vid. págs. 14-16, 18-21, 23-27, 48).

Hyndman, Rob y et al. *forecast: Forecasting functions for time series and linear models*. R package version 8.21. 2023. URL: `https://pkg.robjhyndman.com/forecast/` (vid. pág. 59).

Instituto Nacional de Salud. *Pruebas para la detección molecular de SARS-COV-2 por RT-PCR usadas en Colombia*. Url: `http://www.ins.gov.co/BibliotecaDigital/Pruebas-deteccion-molecular-sars-cov-2-rt-pcr-Colombia.pdf`. 2021 (vid. pág. 55).

Lawrence, Christiano. *Basic Time Series Analysis. Finance 520-1. General Seminar in Finance*. Inf. téc. Kellogg School of Management, Spring, 2011. (Visitado 2011) (vid. pág. 19).

Lehmann, Erich L y George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006 (vid. pág. 44).

McLeod, Allan I y William K Li. “Diagnostic checking ARMA time series models using squared-residual autocorrelations”. En: *Journal of time series analysis* 4.4 (1983), págs. 269-273 (vid. pág. 67).

Moritz, Steffen y Thomas Bartz-Beielstein. “imputeTS: Time Series Missing Value Imputation in R”. En: *The R Journal* 9.1 (2017), págs. 207-218. DOI: 10.32614/RJ-2017-009. URL: https://doi.org/10.32614/RJ-2017-009 (vid. pág. 60).

Niño-Torres, David et al. “Stochastic modeling, analysis, and simulation of the COVID-19 pandemic with explicit behavioral changes in Bogotá: A case study”. En: *Infectious Disease Modelling* 7.1 (2022), págs. 199-211 (vid. pág. 58).

Novales, Alfonso. “Modelos ARCH univariantes y multivariantes”. En: *Departamento de Economıa Cuantitativa. Universidad Complutense de Madrid.(Versión Preliminar). Madrid, Espana* (2013) (vid. págs. 39, 42).

Organizacion, Organizacion Nacional de. *Significado del codigo ISO*. 2023 (vid. pág. 57).

Pascual, Lorenzo, Juan Romo y Esther Ruiz. “Bootstrap prediction for returns and volatilities in GARCH models”. En: *Computational Statistics & Data Analysis* 50.9 (2006), págs. 2293-2312 (vid. pág. 43).

R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/ (vid. pág. 60).

Ríos Gutiérrez, Andrés Sebastián. “Modelos epidemiológicos estocásticos y su inferencia: casos SIS y SEIR”. En: *Departamento de Estadística* (2018) (vid. pág. 14).

Salud, Instituto Nacional de. *Base de datos*. Url: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data. 2023 (vid. págs. 56-58).

Salud, Instituto Nacional De. Url: https://www.ins.gov.co/Noticias/paginas/coronavirus.aspx. 2020 (vid. pág. 12).

Shaphiro, S y MBJB Wilk. “An analysis of variance test for normality”. En: *Biometrika* 52.3 (1965), págs. 591-611 (vid. pág. 33).

Williams, Brandon. “Betreuung: Prof. Dr. Rainer Dahlhaus”. En: (2011) (vid. pág. 44).

A. Conceptos básicos de probabilidad

Para complementar la base teoría de los modelos se utilizan las siguientes definiciones.

Definición A.0.1. Sea $\Omega \neq \emptyset$. Una colección $\Im$ de subconjuntos de Ω es una σ**-álgebra** sobre Ω, si y sólo si
(i) $\Omega \in \Im$
(ii) $A \in \Im$ entonces $A^c \in \Im$
(iii) Si $A_1, A_2, ... \in \Im$ entonces $\bigcup_{i=1}^{\infty} A_i \in \Im$
Los elementos de $\Im$ se llaman **eventos**. La dupla (Ω,$\Im$) se conoce como **espacio medible**.

Es importante establecer una $\sigma - álgebra$ adecuada sobre la cual se pueda definir una variable aleatoria sobre el conjunto de los números reales. Por lo tanto, se define la $\sigma - álgebra$ de Borel.

Definición A.0.2. Sean $\Omega \neq \emptyset$ y $\mathfrak{C}$ una colección de subconjuntos de Ω. Sea

$$\mathcal{M} := \{\Im : \Im \; es \; una \; \sigma - álgebra \; sobre \; \Omega \; tal \; que \; \mathfrak{C} \subseteq \Im\}.$$

La σ-álgebra $\cap_{\Im \in \mathcal{M}} \Im$ se conoce como la σ-**álgebra** generada por $\mathfrak{C}$ y es denotada por $\sigma(\mathfrak{C})$.

Definición A.0.3. La $\sigma - algebra$ sobre $\mathbb{R}$ generada por todos los intervalos de la forma $(-\infty, a]$, con $a \in \mathbb{R}$, se conoce como la σ-álgebra de Borel y se denota por $\mathfrak{B}(\mathbb{R})$.

Ahora daremos la definición de espacio de probabilidad. Esto nos ayudara a definir la medida de probabilidad.

Definición A.0.4. Sean (Ω,$\Im$) un espacio medible y $P : \Omega \to \mathbb{R}$ una función tal que
(i) $P(A) \geq 0$ para todo $A \in \Im$
(ii) $P(\Omega) = 1$

(iii) Si $A_1, A_2, ... \in \Im$ son eventos mutuamente excluyentes, es decir, $A_i \cap A_j \neq \emptyset$ para todo $i \neq j$ entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{+\infty} P(A_i) \tag{A.1}$$

se llama **medida de probabilidad** sobre (Ω,$\Im$). La tripla (Ω,$\Im$,P) se llama **espacio de probabilidad**.

Como un proceso estocástico es un conjunto de variables aleatorias, necesitamos definir variable aleatoria.

Definición A.0.5. Sean $(\Omega, \Im, P)$ un espacio de probabilidad y $(\tilde{\Omega}, \tilde{\Im})$ un espacio medible. Sea X la función tal que

$$X : (\Omega, \Im, P) \to (\tilde{\Omega}, \tilde{\Im})$$

$$\omega \to X(\omega)$$

(i) Se dice que X es una $\Im - \tilde{\Im}$-variable aleatoria si para todo $A \in \tilde{\Im}$ se cumple que $X^{-1}(A) \in \Im$.
(ii) Si $(\tilde{\Omega}, \tilde{\Im})$=$(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ entonces X se conoce como una **variable aleatoria real** sobre $(\Omega, \Im, P)$.

Daremos la definición de distribución de probabilidad de una variable aleatoria, para establecer luego la definición de valor esperado. Con este último concepto, se define la función de media de un proceso estocástico.

Definición A.0.6. Sea $X : \Omega \to \tilde{\Omega}$ una $\Im - \tilde{\Im}$-variable aleatoria con $(\Omega, \Im, P)$ un espacio de probabilidad y $(\tilde{\Omega}, \tilde{\Im})$ un espacio medible. Por notación se escribe

$$X \in B := \{\omega \in \Omega : X(\omega) \in B\}$$

con $B \in \tilde{\Im}$.
La función P_X definida sobre la σ-álgebra $\tilde{\Im}$ por medio de

$$P_X(B) := P(\{X \in B\}) = P(X \in B) \tag{A.2}$$

es una medida de probabilidad sobre $(\tilde{\Omega}, \tilde{\Im})$ llamada la **distribución de la variable aleatoria** X.

La función de distribución caracteriza la distribución de una variable aleatoria, por la cual, se hace la definición.

Definición A.0.7. Sean $(\Omega, \Im, P)$ un espacio de probabilidad y $X : \Omega \to \mathbb{R}$ una variable aleatoria real. La función de distribución de X, denotada por F_X, se define como

$$F_X(x) := P_X((-\infty, x]) = P(X \leq x) \tag{A.3}$$

Dado que no solamente analizaremos una sola variable aleatoria se hace natural definir una forma de estudiar la distribución de probabilidad simultáneamente para varias variables aleatorias. Por tanto, se define a continuación la función de densidad conjunta.

Definición A.0.8. Sea X una variable aleatoria real definida sobre un espacio de probabilidad $(\Omega, \Im, P)$. X es absolutamente continua, si y sólo, si, existe una función real no negativa e integrable f_X tal que

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du, \ \ \forall x \in \mathbb{R},$$

la función f_X se llama la función de densidad de probabilidad de la variable aleatoria X. Si la variable aleatoria X es absolutamente continua, entonces para todo $B \in \mathfrak{B}(\mathbb{R})$ se tiene que $P(X \in B) = \int_B f_X(u)du$.

Definición A.0.9. Sean $X_1, ..., X_n$ variables aleatorias reales definidas todas sobre el espacio de probabilidad $(\Omega, \Im, P)$. Se dice que las variables son **conjuntamente continuas** si existe una función f definida parta todo $(x_1, ..., x_n) \in \mathbb{R}^n$, no negativa e integrable tal que

$$P((X_1, ..., X_n) \in B) = \int_B ... \int_B f(x_1, ..., x_n)dx_1...dx_n$$

para todo $B \in \mathfrak{B}(\mathbb{R}^n)$. La función f se denomina la función de densidad de probabilidad

conjunta.

La distribución normal es importante para nosotros ya que las innovaciones o ruido blanco siguen una distribución normal, por eso se hace necesario definirla .

Definición A.0.10. Se dice que una variable aleatoria X tiene distribución normal de parámetros μ y σ^2, donde $\mu \in \mathbb{R}$ y $\sigma > 0$ si su función de densidad de probabilidad está dada por

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]; \quad x \in \mathbb{R} \tag{A.4}$$

Definición A.0.11. Sean $(\Omega, \Im, P)$ un espacio de probabilidad y $X : \Omega \to \mathbb{R}$ una variable aleatoria real con función de distribución F_X. Si F_X es una función escalonada entonces X es una variable aleatoria de tipo discreto, si F_X es continua se dice que X es de tipo continua y si F_X se expresa como una combinación lineal de una función escalonada y una función entonces se dice que es una variable aleatoria mixta.

Ahora procederemos a definir algunos términos importantes. En primer lugar, la esperanza se refiere al valor promedio de una variable aleatoria. Por otro lado, la varianza es una medida que indica qué tan dispersa está una variable aleatoria de su valor esperado o media. La covarianza determina si hay relación entre dos variables aleatorias, el coeficiente de correlación se utiliza para evaluar que tan relacionadas dos variables aleatorias de manera lineal. Finalmente, dado vamos de definir la covarianza necesitamos definir la probabilidad condicional.

Definición A.0.12. Sean $(\Omega, \Im, P)$ un espacio de probabilidad, A y B elementos de $\Im$ tal que $P(A) > 0$, entonces se define la **probabilidad del evento** B **bajo la condición** A como

$$P(B|A) := \frac{P(A \cap B)}{P(A)}.$$

Un evento B es independiente de un evento A si la ocurrencia del evento A ”no afecta la ocurrencia del evento B”, esto podría interpretarse como $P(B|A) = P(B)$. El

inconveniente de A y de B de esta manera está en que sólo es válida cuando $P(B) > 0$. Por otro parte, si B es independiente de A entonces

$$P(A \cap B) = P(B|A)P(A) = P(B)P(A),$$

razón por la cual se define la independencia de dos eventos y de una familia de eventos como sigue:

Definición A.0.13. Sea $(\Omega, \Im, P)$ un espacio de probabilidad con $A, B \in \Im$. Se dice que A y B son **independientes**, si y sólo si,$P(A \cap B) = P(A)P(B)$. Una familia de eventos $\{A_i : i \in I\}$ se dice que es **independiente**, si y sólo si,

$$P\left(\bigcap_{i \in J} A_j\right) = \prod_{i \in J} P(A_i),\ para\ todo\ subconjunto\ finito\ J \neq \emptyset\ de\ I.$$

Definición A.0.14. Sea X una variable aleatoria real definida sobre el espacio de probabilidad $(\Omega, \Im, P)$.
(i) Si X es una variable aleatoria discreta con valores $x_1, x_2, ...,$ se dice que tiene esperanza si.
$\sum_{i=1}^{+\infty} |x_i| P(X = x_i) < +\infty$, caso en cual la esperanza $\mathbb{E}(X)$ se define como

$$\mathbb{E}(X) = \sum_{i=1}^{+\infty} x_i P(X = x_i) \tag{A.5}$$

(ii) Si $\mathbb{E}(X^2)$ y $\mathbb{E}(X)$ existen se define la varianza como

$$\mathbb{V}(X) = \mathbb{E}([X - \mathbb{E}(X)]^2) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \tag{A.6}$$

(iii) Sean X y Y variables aleatorias sobre $(\Omega, \Im, P)$ tales que $\mathbb{E}(X^2) < +\infty$ y $\mathbb{E}(Y^2) < +\infty$. Se define la covarianza de X y Y como

$$Cov(X, Y) = \mathbb{E}([X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{A.7}$$

(iv) Sean X y Y variables aleatorias sobre (Ω, F, P) tales que $\mathbb{E}(X^2) < +\infty$ y $\mathbb{E}(Y^2) < +\infty$.

Se define el coeficiente de correlación de X y Y como

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}} \tag{A.8}$$

La función indicadora nos ayudara para posteriormente definir la esperanza condicional, ya que nos permite calcular la media esperada de una variable aleatoria en función de la información disponible.

Definición A.0.15. Sean $(\Omega, \Im, P)$ un espacio de probabilidad y $B \in \Im$. Se define la función indicadora de B, para todo $\omega \in \Omega$, como

$$1_B(\omega) = \begin{cases} 1 & si\ \omega \in B \\ \\ 0 & en\ otro\ caso \end{cases}$$

Definición A.0.16. Sean X una variable aleatoria real sobre el espacio de probabilidad $(\Omega, \Im, P)$,
Si $B \in \Im$ con $P(B) > 0$ y $\mathbb{E}(X1_B(\omega)) < +\infty$, se define la esperanza condicional de X dado B como

$$\mathbb{E}(X|B) := \frac{\mathbb{E}(X1_B)}{P(B)} \tag{A.9}$$

Definición A.0.17. Si $\mathfrak{G}$ es una sub-σ-álgebra de $\Im$ se define la esperanza condicional de X, dada $\mathfrak{G}$ como una variable aleatoria $\mathfrak{G}$-medible, (es decir $\mathbb{E}(X|1_A < +\infty), \forall A \in \mathfrak{G}$) denotada por $\mathbb{E}(X|\mathfrak{G})$, tal que

$$\mathbb{E}([X - \mathbb{E}(X|\mathfrak{G})]1_G) = 0,\ para\ todo\ G \in \mathfrak{G} \tag{A.10}$$

donde $E(X|\mathfrak{G})$ corresponde a la esperanza condicional.

B. Conceptos de epidemiología

Definición B.0.1. Virus. Un virus es un microorganismo infeccioso que consta de un segmento de ácido nucleico (ADN o ARN)rodeado por una cubierta protica.Un virus no puede replicarse solo; por el contrario, debe infectar a las células y usar componentes de la célula huésped para fabricar copias de sí mismo.

Definición B.0.2. Pandemia. Una epidemia se produce cuando una enfermedad contagiosa se propaga rápidamente en una población determinada, afectando simultáneamente a un gran número de personas durante un periodo de tiempo concreto.

Definición B.0.3. Incidencia. Número de casos nuevos de una enfermedad que se diagnostican.

Definición B.0.4. Síntoma. Problema físico o mental que presenta una persona, el cual puede indicar una enfermedad o afección. Los síntomas no se pueden observar y no se manifiestan en exámenes médicos. Algunos ejemplos de síntomas son el dolor de cabeza, el cansancio crónico, las náuseas y el dolor.

Definición B.0.5. Enfermedad. La OMS define enfermedad como .Alteración o desviación del estado fisiológico en una o varias partes del cuerpo, por causas en general conocidas, manifestada por síntomas y signos característicos, cuya evolución es más o menos previsible.

Definición B.0.6. Enfermedad infecciosa. Las enfermedades infecciosas son trastornos causados por organismos, como bacterias, virus, hongos o parásitos. Muchos organismos viven dentro y fuera de nuestros cuerpos. Normalmente son inofensivos o incluso útiles. Pero bajo ciertas condiciones, algunos organismos pueden causar enfermedades.

Definición B.0.7. Número reproductivo básico. El número reproductivo básico de una pandemia, también conocido como R0, es una medida epidemiológica que representa el promedio de nuevos casos que un individuo infectado generaría en una población susceptible