



Trabajo Práctico N°2

Fundamentos de Análisis de Datos
Maestría en Ciencia de Datos

Estudiantes:

Ing. Flores, Matías Gabriel
matflores@itba.edu.ar

Ing. Loiseau, Matías
mloiseau@itba.edu.ar

Docente:

Dra. Rey, Andrea A.

Septiembre 2023

Contents

1	Ejercicio N° 1	3
1.1	Primer punto	3
1.2	Segundo punto	3
1.3	Tercer punto	5
2	Ejercicio N° 2	9
2.1	Primer punto	9
2.2	Segundo punto	10
2.3	Tercer punto	10
3	Ejercicio N° 3	10
3.1	Primer punto	11
3.2	Segundo punto	11
3.3	Tercer punto	11
3.4	Cuarto punto	12
3.5	Quinto punto	12
4	Ejercicio N° 4	13
4.1	Primer punto	13
4.2	Segundo punto	13
4.3	Tercer punto	16
4.4	Cuarto punto	18
4.5	Quinto punto	20
4.6	Sexto punto	20

1 Ejercicio N° 1

Vamos a trabajar con el archivo *MedidasCorporales.xlsx*.

1.1 Primer punto

Consigna: ¿Cuántos registros y cuántas variables tiene el conjunto de datos? ¿Todas las variables son numéricas?

El set de datos de Medias Corporales tiene 507 registros y 24 variables. Verificamos que no haya nulos. Esto se puede visualizar en la Tabla 1. Además, durante el análisis de los datos pudimos verificar que todas las variables son numéricas.

	Valor
Registros	507
Columnas	24
Nulos	0

Table 1: Cantidad de campos en el set de datos.

1.2 Segundo punto

Consigna: Regresión lineal simple: ¿cómo influye la altura en el peso?

- Realizar una regresión lineal simple y escribir el modelo teórico resultante.
- ¿Cuáles son las estimaciones de la ordenada al origen y de la pendiente? ¿Son estos coeficientes de regresión significativos?
- Calcular el error estándar residual, el coeficiente de determinación R^2 y su valor ajustado. ¿Qué se podría concluir sobre la bondad de ajuste del modelo?

Teniendo en cuenta la consigna, en primer lugar analizamos como se distribuye la variable de peso en función a la de altura mediante un diagrama de dispersión, figura 1.1.

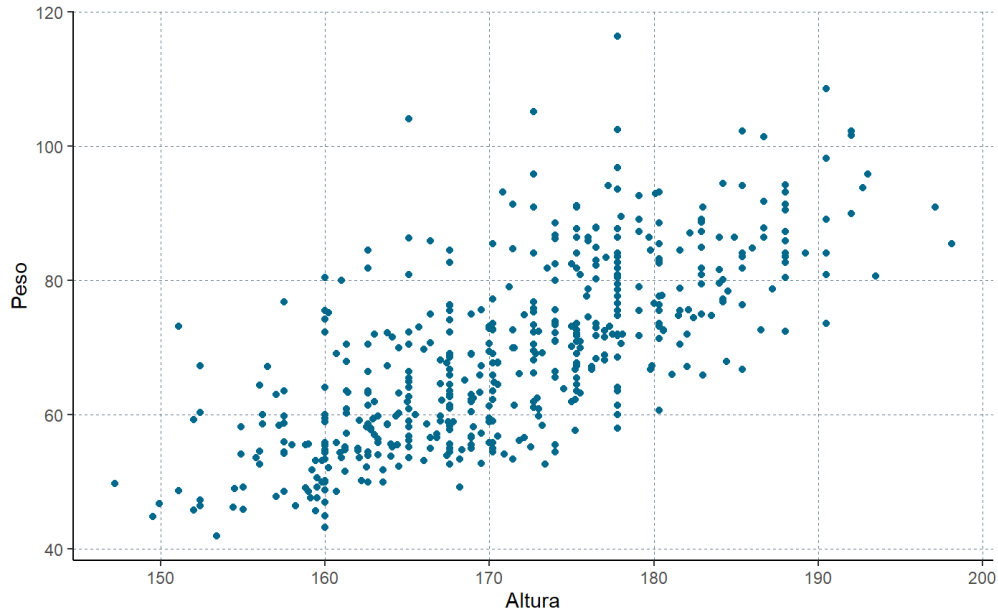


Figure 1.1: Diagrama de dispersión entre Peso y Altura.

Analizando el gráfico obtenido, podemos ver que pareciera haber una correlación positiva entre ambas variables. Realizando el calculo de correlación obtenemos como resultado un valor de 0.717, por lo tanto podemos determinar que se encuentran bastante correlacionadas.

Luego, procedimos a realizar el modelo de regresión lineal, teniendo como variable dependiente al peso, y variable independiente a la altura. Como resultado, el modelo teórico queda de la siguiente forma:

$$\hat{y} = -105.011 + 1.018x \quad (1)$$

donde x representa la altura e \hat{y} representa el peso estimado por el modelo.

Graficando esta recta de regresión lineal sobre el diagrama de la figura 1.1, nos queda el grafico de la figura 1.2, donde podemos observar que la linea se ajusta a los puntos.

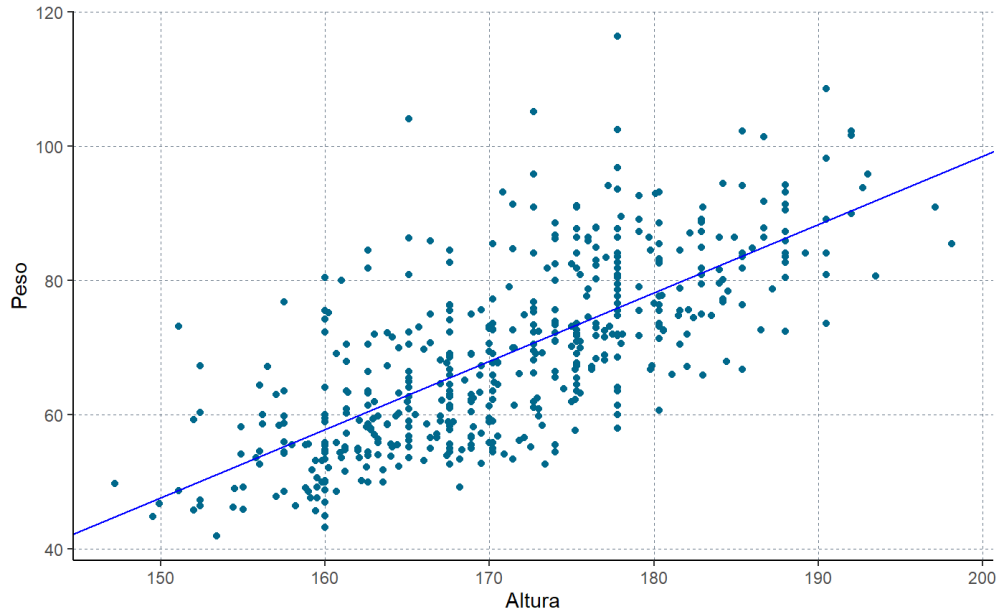


Figure 1.2: Diagrama de dispersión entre Peso y Altura con Recta.

La estimación de la ordenada al origen es -105.011 y de la pendiente es 1.018 . Al ser una pendiente positiva, a medida que el valor de x aumente, el de y también tenderá a aumentar. Estos coeficientes de regresión tienen un alto grado de significancia según los resultados obtenidos.

El error estándar residual resultante determina que la estimación promedio del peso puede diferir en 9.308kg . En cuanto al coeficiente de determinación R^2 obtenido, consideramos que es un valor significativamente moderado. Por ultimo, el valor ajustado del coeficiente nos dio bastante similar al original. Estos valores se pueden observar en la Tabla 2.

Variable	Valor
Residuo	9.308
R^2	0.5145
Valor ajustado	0.5136
Estadístico-F	354.4
P-Valor	$< 2.2e^{-16}$

Table 2: Parámetros de salida del modelo.

A partir de los resultados obtenidos, determinamos que el 51% de las variaciones de los pesos están explicadas por la altura de las personas. Además, podemos concluir que existe una relación lineal significativa entre la altura y el peso, ya que el estadístico-f observado tiene un valor alto con un p-valor menor a 0.05 , rechazando la hipótesis nula.

1.3 Tercer punto

Consigna: Regresión lineal múltiple: ¿cómo influyen las medidas consideradas en el peso?

- Guardar en la variable n la cantidad total de registros.
- Fijar una semilla igual a 1234 y correr el siguiente comando:

```
muestras <- 1:n %>%
  createDataPartition(p=0.8, list=FALSE)
```

Usar la variable muestras para separar aleatoriamente el conjunto de registros en conjuntos de entrenamiento y de prueba. ¿Qué porcentaje de los datos integra cada uno de estos conjuntos?

- (c) A partir del conjunto de entrenamiento, realizar el modelo de regresión lineal múltiple con todas las variables involucradas.
- (d) A partir del conjunto de entrenamiento, realizar el modelo de regresión lineal múltiple con las variables que presenten un nivel de confianza de al menos el 95%.
- (e) Utilizando el conjunto de prueba, calcular el error cuadrático medio:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

donde n es la cantidad total de predicciones, y_i es el valor real de la variable respuesta e \hat{y}_i es el valor predicho. ¿Cuál de los dos modelos muestra un valor menor de MSE?

Para este ejercicio, en primer lugar, dividimos el dataset en un conjunto de entrenamiento y uno de prueba, mediante la semilla y el comando establecidos. Una vez realizado esto, el set de datos de entrenamiento resultante nos quedó con un 80% del data set original, mientras que el de prueba con un 20%.

Luego de haber separado el dataset, pasamos a realizar los modelos de regresión lineal múltiple solicitados. En este caso, realizamos 2 modelos para ver que como influyen las variables en el peso. En el primer modelo utilizamos todas las variables del dataset, mientras que para el segundo utilizamos solo las variables que presentaran al menos un 95

Al realizar el primer modelo, obtuvimos los coeficientes de regresión que se encuentran en la Tabla 3.

Variable	Coficiente
Ordenada al Origen	-116.74490
Biacromial	-0.10649
Ancho pélvico	0.13638
Bitrocantérico	-0.05228
Profundidad de pecho	0.29965
Diámetro de pecho	0.14243
Diámetro de codo	0.14037
Diámetro de muñeca	0.30880
Diámetro de rodilla	0.36165
Diámetro de tobillo	0.29943
Contorno de hombro	0.05855
Contorno de pecho	0.18343
Contorno de cintura	0.31672
Contorno de ombligo	0.01888
Contorno de cadera	0.26093
Contorno de muslo	0.23083
Contorno de bíceps	0.02407
Contorno de antebrazo	0.43032
Contorno de rodilla	0.27090
Contorno de pantorrilla	0.35148
Contorno de tobillo	-0.06504
Contorno de muñeca	-0.15169
Edad	-0.08307
Altura	0.27063

Table 3: Tabla con los Coeficientes del Modelo 1.

Luego, procedimos a analizar el nivel de confianza obtenido, que se encuentra detallado en la Tabla 4.

Variable	2.5	97.5
Ordenada al Origen	$-1.22e^{+2}$	-111.29
Biacromial	$-2.49e^{-1}$	0.036
Ancho pélvico	$-4.03e^{-3}$	0.276
Bitrocantérico	$-2.47e^{-1}$	0.14
Profundidad de pecho	$-1.54e^{-1}$	0.44
Diámetro de pecho	$-3.06e^{-2}$	0.315
Diámetro de codo	$-2.65e^{-1}$	0.54
Diámetro de muñeca	$-1.78e^{-1}$	0.79
Diámetro de rodilla	$4.4e^{-2}$	0.67
Diámetro de tobillo	$-4.71e^{-2}$	0.64
Contorno de hombro	$-9.94e^{-3}$	0.12
Contorno de pecho	$9.75e^{-2}$	0.26
Contorno de cintura	$2.62e^{-1}$	0.37
Contorno de ombligo	$-3.43e^{-2}$	0.07
Contorno de cadera	$1.58e^{-1}$	0.36
Contorno de muslo	$1.19e^{-1}$	0.34
Contorno de bíceps	$-1.52e^{-1}$	0.20
Contorno de antebrazo	$1.35e^{-1}$	0.72
Contorno de rodilla	$9.89e^{-2}$	0.44
Contorno de pantorrilla	$2.02e^{-1}$	0.50
Contorno de tobillo	$-2.77e^{-1}$	0.14
Contorno de muñeca	$-5.95e^{-1}$	0.29
Edad	$-1.09e^{-1}$	-0.056
Altura	$2.31e^{-1}$	0.309

Table 4: Tabla con los Niveles de Confianza de las Variables.

A partir de estos valores, determinamos las variables que utilizaríamos en el modelo 2. A partir de la confianza calculada, tomamos en cuenta todos los valores que no pasaran por cero. Por lo tanto, las variables que utilizamos fueron "Profundidad de pecho", "Diámetro de rodilla", "Contorno de pecho", "Contorno de cintura", "Contorno de cadera", "Contorno de muslo", "Contorno de antebrazo", "Contorno de rodilla", "Contorno de pantorrilla", "Edad" y "Altura". Los coeficientes obtenidos para este modelo se encuentran en la siguiente Tabla 5.

Variable	Coeficiente
Ordenada al Origen	-117.85
Profundidad de pecho	0.27092
Diámetro de rodilla	0.56093
Contorno de pecho	0.25826
Contorno de cintura	0.32925
Contorno de cadera	0.28963
Contorno de muslo	0.22904
Contorno de antebrazo	0.47994
Contorno de rodilla	0.23341
Contorno de pantorrilla	0.34611
Edad	-0.07167
Altura	0.28984

Table 5: Tabla con los Coeficientes del Modelo 2.

Finalmente, habiendo entrenado los dos modelos sobre el conjunto de entrenamiento, procedimos a realizar las predicciones sobre el conjunto de prueba. Una vez obtenidas las predicciones, llevamos a cabo el calculo del error cuadrático medio para cada uno. Como resultado, obtuvimos los valores que se detallan en la Tabla 6.

Modelo	Residuo	R^2	R^2 Ajustado	Estadístico F	P-Valor	MSE
1	2.077	0.97	0.97	710	$< 2.2e^{-16}$	5.630
2	2.098	0.97	0.97	1454	$< 2.2e^{-16}$	5.369

Table 6: Errores cuadráticos medios.

A partir de los resultados obtenidos, podemos observar que el Estadístico F del modelo 2 tiene mas que el doble de tamaño que el del primer modelo, determinando que posee una mejor capacidad de explicar la variabilidad del peso en base a las variables involucradas. Los otros resultados dieron bastante similares para ambos, siendo el residuo levemente mas alto en el modelo 2. Por ultimo, el segundo modelo es el que presenta un menor error cuadrático medio, siendo el que tiene un mejor ajuste, y el mas eficiente de los dos.

2 Ejercicio N° 2

Vamos a trabajar con el archivo *Dolor.xlsx*. El mismo contiene una muestra de 3504 pacientes que acudieron a un centro de salud presentando dolor en el pecho. Para estos pacientes, se recogieron diversas medidas. En el caso de las variables estrechamiento de arterias coronarias y de tres arterias coronarias, ambas son variables binarias que indican la presencia de estrechamiento en alguna de las arterias coronarias de al menos un 75% (valor igual a 1) o no (valor igual a 0). En cuanto a la variable sexo, 0 corresponde a masculino y 1 a femenino.

2.1 Primer punto

Consigna: Realizar un modelo de regresión logística simple, que estudie la presencia de estrechamiento en alguna arteria coronaria explicada por el colesterol. Escribir la ecuación del modelo resultante y calcular la probabilidad de que una persona con un nivel de colesterol igual a 199 presente estrechamiento arterial.

En primer lugar, realizamos un analisis del dataset con el que ibamos a trabajar. Este conjunto de datos se encuentra compuesto por 7 variables numericas. Lo primero que pudimos encontrar, fue la presencia de valores nulos, por lo cual decidimos eliminarlos. Una vez realizado esto, detectamos que la variable "Colesterol", si bien estaba compuesta por valores numericos, era de tipo "Char" con campos llamados "NA", lo cual sería un conflicto al realizar el modelo. Por lo tanto, eliminamos estos valores y convertimos la variable a tipo numérica, quedando así con 2258 registros.

Para realizar el modelo de regresión logística simple, decidimos utilizar la variable "Estrechamiento arterias coronarias" como nuestra variable dependiente, y "Colesterol" como la independiente. Mediante este modelo, obtuvimos los coeficientes necesarios para determinar la ecuación resultante.

$$p(x) = \frac{\exp(-0.7525280 + 0.0062268x)}{1 + \exp(-0.7525280 + 0.0062268x)} \quad (2)$$

Mediante el modelo implementado, pudimos calcular la probabilidad de que una persona con un nivel de colesterol de 199 presente estrechamiento arterial, siendo este un 61

$$p(199) = 0.6193053 \quad (3)$$

2.2 Segundo punto

Consigna: Realizar un modelo de regresión logística múltiple, que estudie la presencia de estrechamiento en alguna arteria coronaria usando todas las variables no categóricas como variables explicativas. ¿Qué puede decirse sobre la significancia de las variables predictoras?

Las variables que tuvimos en cuenta para este punto fueron "Edad", "Días con Síntomas" y "Colesterol". Al realizar el modelo, obtuvimos los resultados que se encuentran en la Tabla 7.

Variable	Estimado	Error	z value	P-Valor	Código de Significancia
(Intercept)	-3.38	0.347923	-9.729	$< 2e^{-16}$	***
Edad	0.052507	0.005317	9.876	$< 2e^{-16}$	***
Días con síntomas	-0.001007	0.000916	-1.099	0.272	
Colesterol	0.006394	0.000976	6.551	$5.72e^{-11}$	***

Table 7: Resultados del Modelo de Regresión Logística.

Mediante estos resultados, podemos ver que tanto "Edad" como "Colesterol", tienen un coeficiente positivo, por lo que el aumento de estas dos variables se puede aumentar las probabilidades de tener estrechamiento en alguna arteria coronaria, teniendo una fuerte significancia en el modelo. Por otro lado, la cantidad de Días con Síntomas no parece tener una relación significativa con la variable dependiente.

2.3 Tercer punto

Consigna: Replicar el modelo anterior pero diferenciando entre mujeres y varones. ¿Existen diferencias entre las significancias de las variables explicativas en función del sexo? Justificar la respuesta.

Realizando el anterior modelo para mujeres y varones por separado, a simple vista, no parece tener un fuerte impacto en los valores obtenidos para cada uno. Los valores obtenidos para cada uno se encuentran en la Tabla 8.

Variable	Estimado	Error	z value	P-Valor	Código de Significancia
(Intercept) Hombre	-5.2804362	0.5301514	-9.960	$< 2e^{-16}$	***
Edad Hombre	0.0862912	0.0079183	10.898	$< 2e^{-16}$	***
Días con síntomas Hombre	-0.0005885	0.0014400	-0.409	0.683	
Colesterol Hombre	0.0106062	0.0015173	6.990	$2.75e^{-12}$	***
(Intercept) Mujer	-4.885106	0.611157	-7.993	$1.31e^{-15}$	***
Edad Mujer	0.049992	0.009851	5.075	$3.87e^{-07}$	***
Días con síntomas Mujer	-0.003019	0.001545	-1.954	0.0507	.
Colesterol Mujer	0.008202	0.001545	5.309	$1.1e^{-07}$	***

Table 8: Resultado para Conjunto de Hombres y Conjunto de Mujeres.

Para el grupo de Mujeres, la variable de Días con Sintomas pareciera tener una mayor significancia, teniendo un p-valor cerca al 0.05, a diferencia del grupo de Hombres. Por otro lado, las otras variables parecen tener la misma significancia en ambos grupos.

3 Ejercicio N° 3

Vamos a trabajar con el archivo Europa.xlsx.

3.1 Primer punto

Consigna: ¿Cuáles son las variables de interés?

Las variables de interés del set de datos *Europa* son:

- Área
- PBI
- Inflación
- Expectativa de vida
- Población militar
- Crecimiento de la población
- Tasa de desempleo

Hay que tener en cuenta que en el mismo archivo se encuentra el campo País, pero este sirve como índice y no como variable.

3.2 Segundo punto

Consigna: Calcular la matriz de covarianza de los datos y analizar si es inversible.

Primero realizamos la matriz de covarianzas y luego aplicamos la función para determinar si era inversible. La matriz se encuentra en la Tabla 9, donde 1 es “Área”, 2 es “PBI”, 3 es “Inflación”, 4 es “Expectativa de Vida”, 5 es “Población Militar”, 6 es “Crecimiento de Población” y 7 es “Tasa de Desempleo”.

	1	2	3	4	5	6	7
1	2.74e+10	-3.32e+08	7.40e+04	-1.14e+04	1.34e+04	-7363.91	19707.05
2	-3.32e+08	2.10e+08	-9.99e+03	3.24e+04	-3.30e+03	5535.55	-35802.46
3	7.40e+04	-9.99e+03	1.95e+00	-3.02e+00	5.41e-02	-0.33	1.30
4	-1.14e+04	3.24e+04	-3.02e+00	1.01e+01	-1.61e-01	1.23	-3.66
5	1.34e+04	-3.30e+03	5.41e-02	-1.61e-01	6.42e-01	-0.11	1.09
6	-7.36e+03	5.53e+03	-3.36e-01	1.23e+00	-1.13e-01	0.25	-0.41
7	1.97e+04	-3.58e+04	1.30e+00	-3.66e+00	1.09e+00	-0.41	21.88

Table 9: Matriz de Covarianzas.

El resultado de la invarianza fue 1.235391e+19. Como dicho valor es distinto a 0, determinamos que es inversible.

3.3 Tercer punto

Consigna: ¿Cuál es el mayor autovalor de la matriz de covarianzas?

A partir de los autovalores obtenidos en la matriz de covarianzas, determinamos que el mayor autovalor es el de la variable “Área”

3.4 Cuarto punto

Consigna: Realizar un PCA y hallar la cantidad necesaria de componentes principales para explicar al menos el 90% de la varianza total de los datos.

Variable Descriptiva	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.796	1.0896	1.0311	0.8777	0.67665	0.4107	0.35446
Proportion of Variance	0.461	0.1696	0.1519	0.1100	0.06541	0.0241	0.01795
Cumulative Proportion	0.461	0.6306	0.7825	0.8925	0.95795	0.9820	1.00000

Table 10: Variables descriptivas de los componentes principales.

En base a los componentes principales calculados, y como se puede observar en la tabla 10 determinamos que con 4 componentes podemos explicar un 89.25% de la varianza total de los datos, llegando así casi un 90% de ellos.

3.5 Quinto punto

Consigna: Realizar e interpretar un gráfico que visualice la contribución de las variables en las dos primeras componentes principales.

Teniendo en cuenta las funciones realizadas en los puntos anteriores, procedemos a realizar un grafico para demostrar la contribución de las variables en las dos primeras componentes principales.

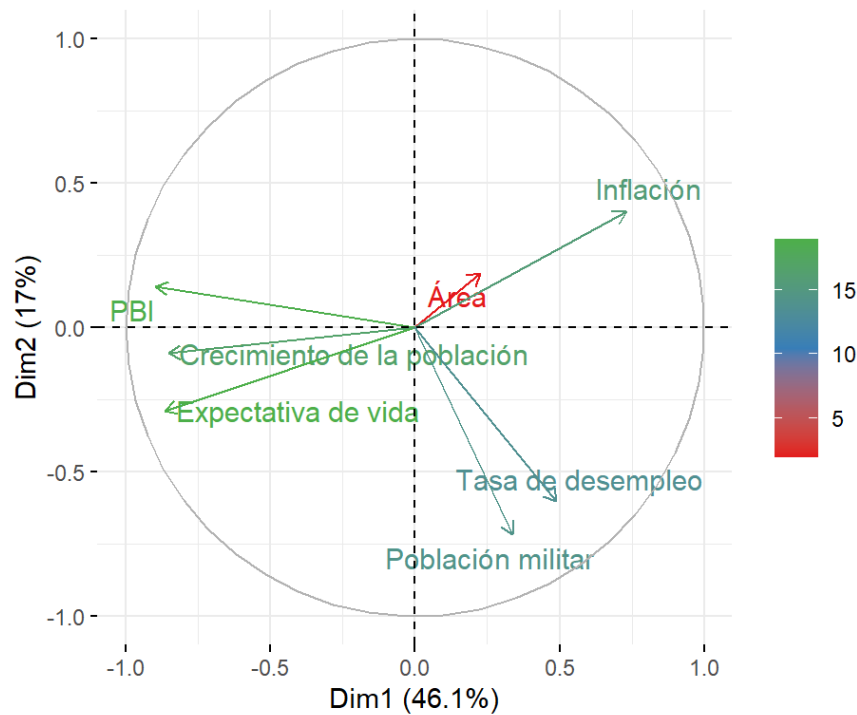


Figure 3.1: Representación y comparación de las direcciones de las variable en función de las dimensiones 1 y 2.

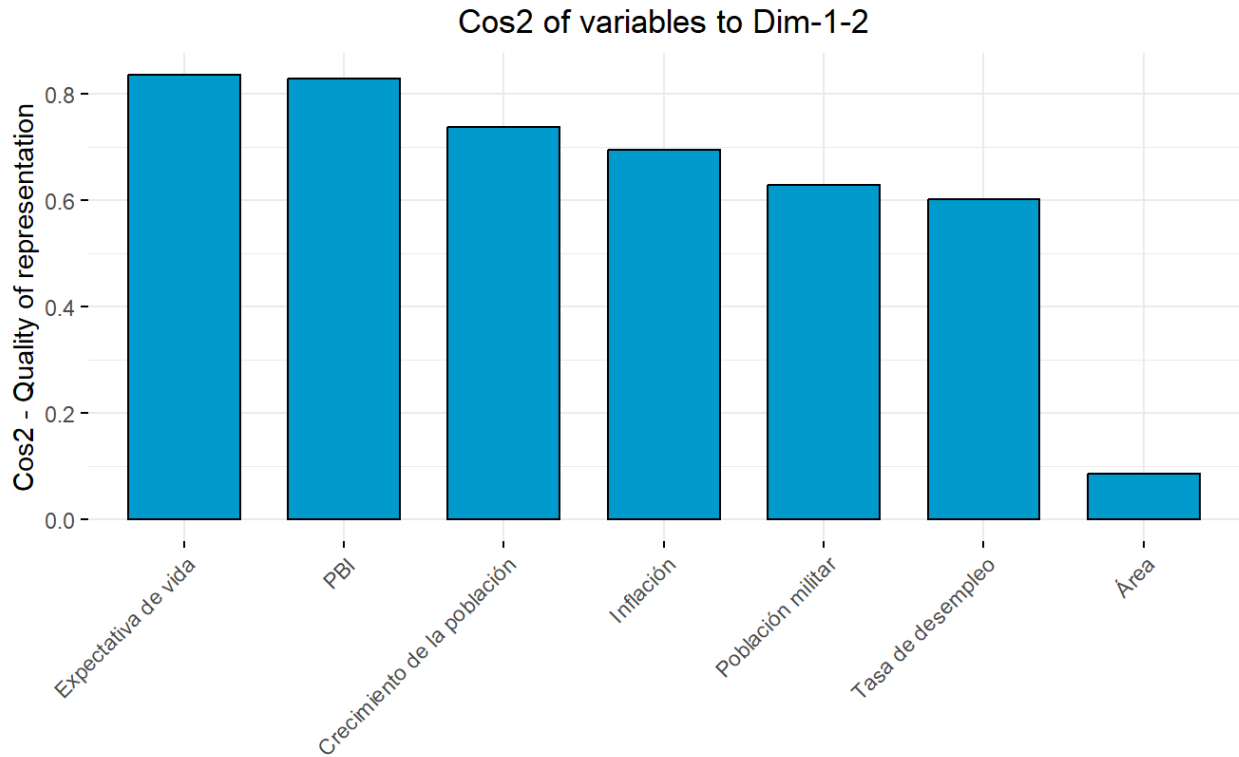


Figure 3.2: Calidad de la representación de las dimensiones 1 y 2.

En la figura 3.1 observamos la calidad de representación de las variables y a su vez podemos comparar una con otras para ver que tan parecidas son entre ellas. En la otra figura 3.2 representa la misma información a excepción de la dirección de las variables.

4 Ejercicio N° 4

Vamos a considerar el conjunto de datos JohnsonJohnson disponible en R.

4.1 Primer punto

Consigna: ¿Qué tipo de datos mide esta serie de tiempo? ¿Cuál es el período de tiempo analizado?

Esta serie de tiempo contiene las ganancias en dolares trimestrales de "Johnson y Johnson", desde el año 1960 hasta 1980.

4.2 Segundo punto

Consigna: Graficar la serie tiempo, junto con sus descomposiciones aditiva y multiplicativa. ¿Se observa tendencia? ¿Se observa estacionalidad?

Lo primero que realizamos fue el gráfico de la serie de tiempo. Esta se puede ver en la Figura 4.1.

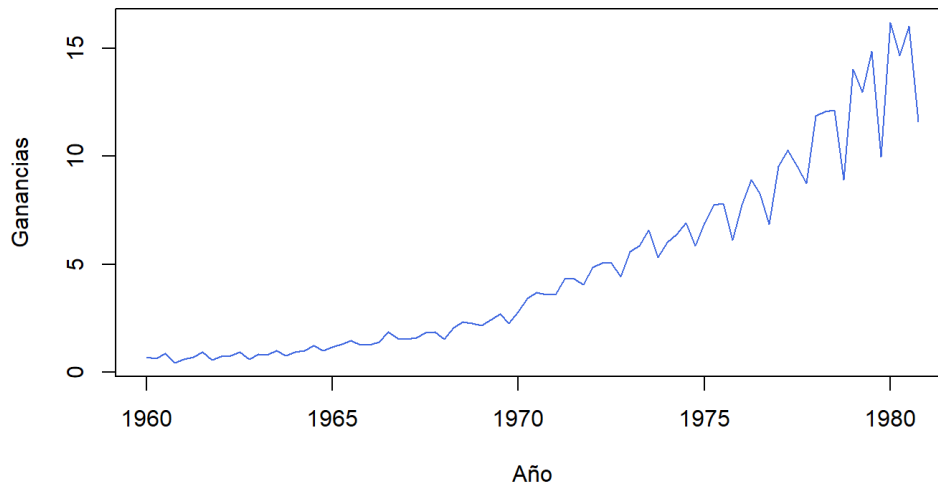


Figure 4.1: Representación de la Serie de Tiempo.

En este gráfico podemos visualizar un incremento en las ganancias a lo largo del tiempo. Para analizar la tendencia y la estacionalidad en detalle, vamos a realizar las descomposiciones aditiva y multiplicativa.

Primero vamos a analizar la descomposición aditiva. La tendencia se puede visualizar en la Figura 4.2.

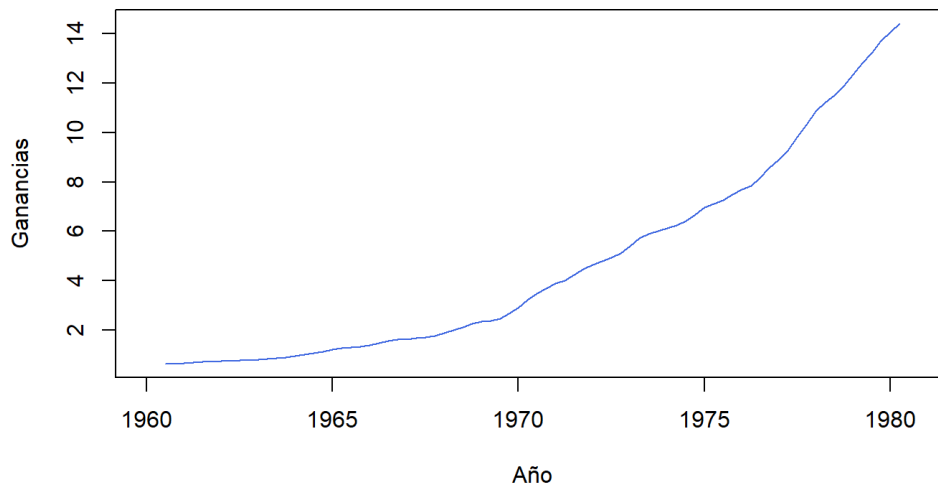


Figure 4.2: Tendencia Aditiva.

A partir del gráfico, vemos que puede llegar a tener una tendencia positiva. Para saber con exactitud el valor de tendencia, realizamos el calculo de la fuerza de tendencia. Al realizar el calculo nos da como resultado un 0.97. Por lo que podemos determinar que existe una fuerte tendencia positiva.

Lo siguiente que se analizó fue la estacionalidad, que se puede visualizar en la Figura 4.3.

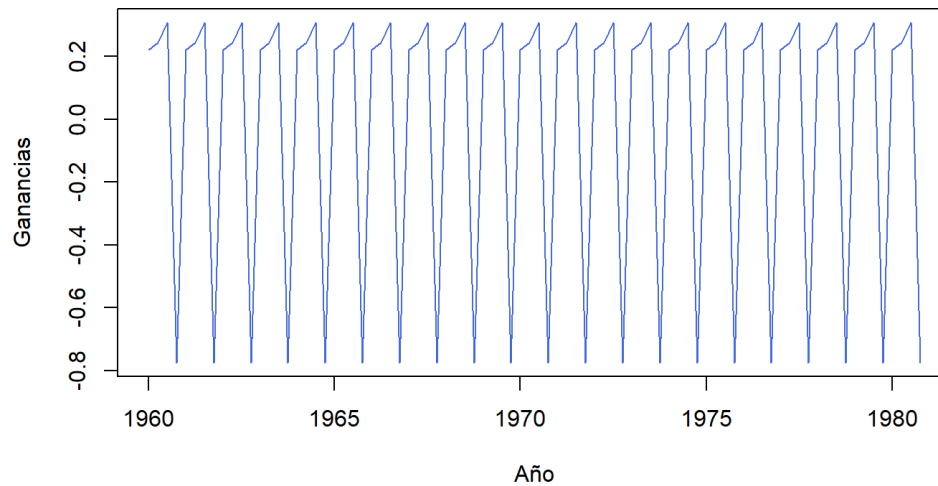


Figure 4.3: Estacionalidad Aditiva.

Calculamos la fuerza de la estacionalidad, la cual nos da como resultado 0.32, siendo un resultado relativamente chico.

Luego, pasamos a analizar la descomposición multiplicativa. El grafico de tendencia se puede ver en la Figura 4.4.

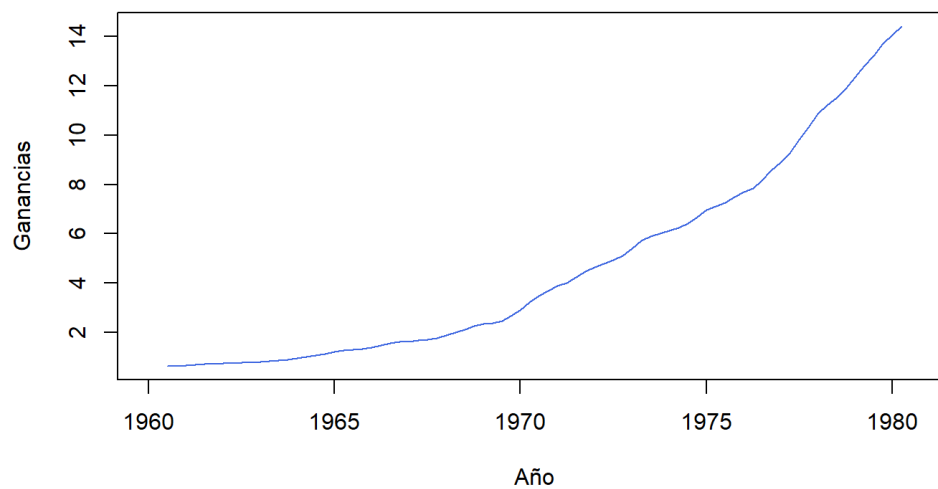


Figure 4.4: Tendencia Multiplicativa.

Nuevamente, parece tener una tendencia alta. Al realizar el calculo de fuerza, nos da como resultado 0.99,

por lo tanto, determinamos que tiene una alta tendencia.

Finalmente, realizamos el análisis de la estacionalidad, representado en la Figura 4.5.

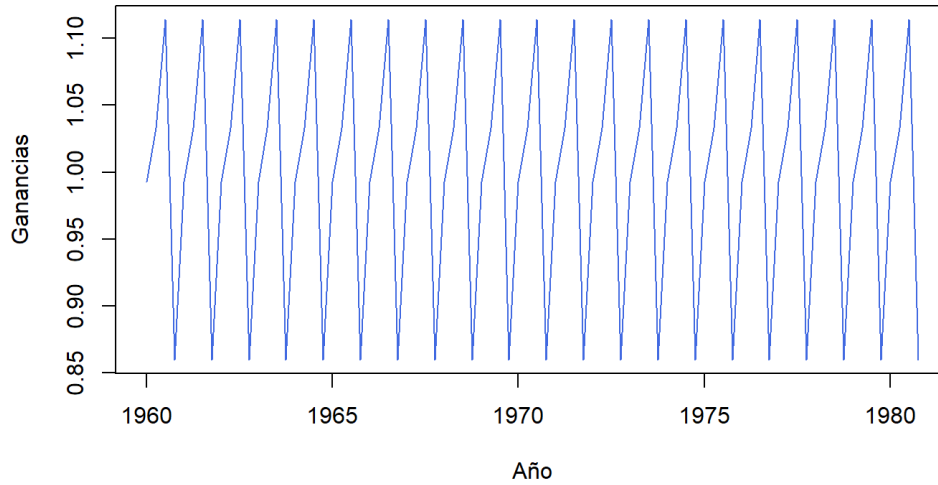


Figure 4.5: Estacionalidad Multiplicativa.

Nuevamente, realizando el calculo de fuerza de estacionalidad, nos da como resultado 0.54. Dicho valor es bastante superior al 0.32 de la aditiva.

En base a los análisis realizados en este punto, podemos determinar que existe una alta tendencia en la serie de tiempo, por otro lado, los valores de estacionalidad no son significativamente altos, aunque tampoco podemos determinar que sean despreciables.

4.3 Tercer punto

Consigna: Analizar la conveniencia de aplicar la transformación de Box-Cox.

Teniendo en cuenta el análisis realizado en el punto anterior, y revisando los residuos de las descomposiciones realizadas, vamos a proceder a verificar si es conveniente realizar una transformación de BoxCox sobre la serie de tiempo. Mediante esta transformación se estacionaría la serie de tiempo, lo cual nos va a ayudar a reducir la tendencia y estacionalidad que tiene actualmente, para más adelante poder realizar la implementación de un modelo ARIMA.

Lo primero que hicimos fue comprobar mediante los test de Phillips-Perron y de Kwiatkowski-Phillips-Schmidt-Shin si la serie de tiempo original es estacionaria. A partir de los resultados obtenidos, determinamos que no lo es, por lo cual procedemos a realizar la transformación de BoxCox. Para esto, realizamos la búsqueda del mejor parámetro λ . En nuestro caso, nos dio 0.15. Teniendo este valor, procedemos a realizar la transformación. En la Figura 4.6 se puede ver la comparativa entre la serie original y la transformada.

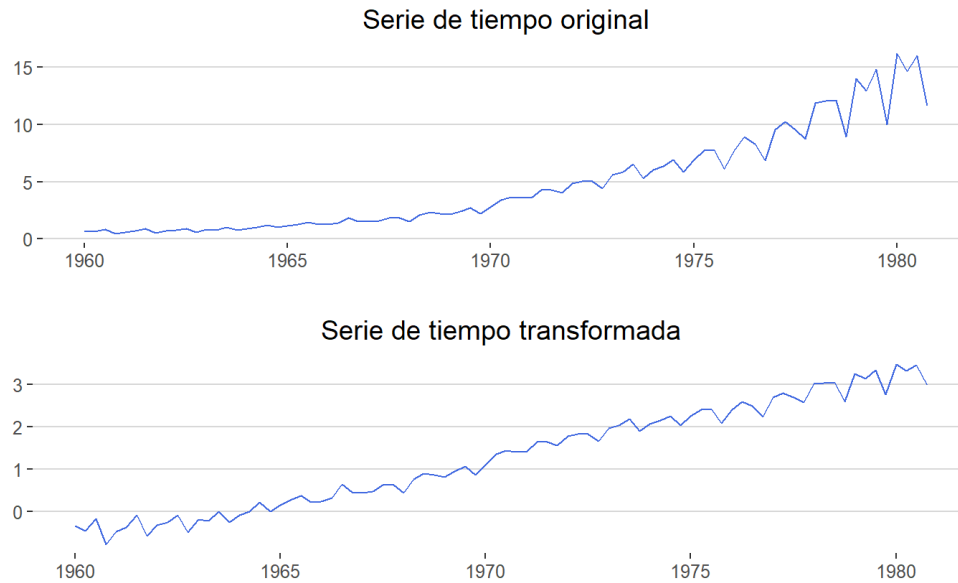


Figure 4.6: Serie Original vs Transformada.

Analizando los gráficos, vemos que pareciera haber una reducción en la tendencia y estacionalidad. Para comprobar si es estacionaria, realizamos nuevamente las pruebas anteriormente mencionadas. Una vez hecho esto, obtenemos que sigue sin ser estacionaria. Por lo cual decidimos realizar una diferenciación sobre ella para eliminar la tendencia que queda. El resultado de esto se puede apreciar en la Figura 4.7.

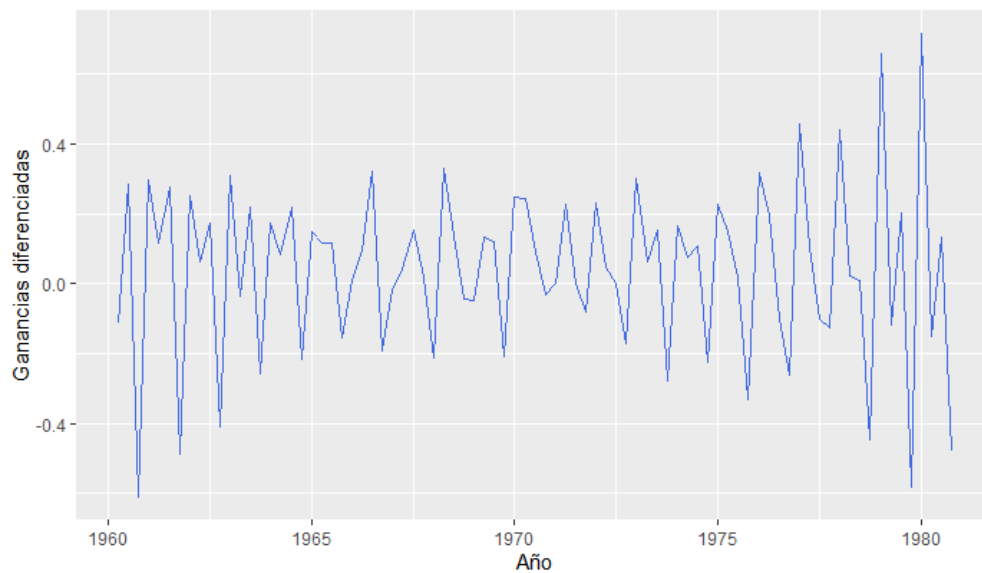


Figure 4.7: Serie Diferenciada.

Realizamos devuelta los test de Phillips-Perron y de Kwiatkowski-Phillips-Schmidt-Shin, para determinar si la serie resultante es estacionara. Los resultados obtenidos para ambas pruebas nos dan p valores para determinar que la serie resultante es estacionaria. Por lo tanto, decidimos conveniente realizar la transformación

de Box-Cox, además de una diferenciación de nivel 1.

4.4 Cuarto punto

Consigna: Usar toda la información de todos los años salvo los dos últimos para realizar un modelo ARIMA automático y uno personalizado, explicando la elección de los órdenes elegidos y teniendo en cuenta lo concluido en el punto anterior. Trabajar con $1 \leq p \leq 14$ y $1 \leq q \leq 30$.

Teniendo en cuenta los resultados obtenidos anteriormente, trabajamos con la serie transformada y diferenciada. Lo primero que hicimos fue realizar el modelo de ARIMA personalizado. Para esto, dividimos el la serie diferenciada en dos conjuntos, uno de entrenamiento con los datos de los años 1960 a 1978, y otro de prueba solamente con el ultimo año (1980) para lo que se solicita en el sexto punto.

A partir de esta división, procedimos a determinar el mejor valor de p y q para utilizar en la implementación del modelo. Realizamos los gráficos para la función de autocorrelación parcial (PACF), mediante el cual podemos observar el p con menor valor; y el de la función de autocorrelación (ACF), para determinar el q menor. Estos gráficos se pueden visualizar en la Figura 4.8 y Figura 4.9, respectivamente.

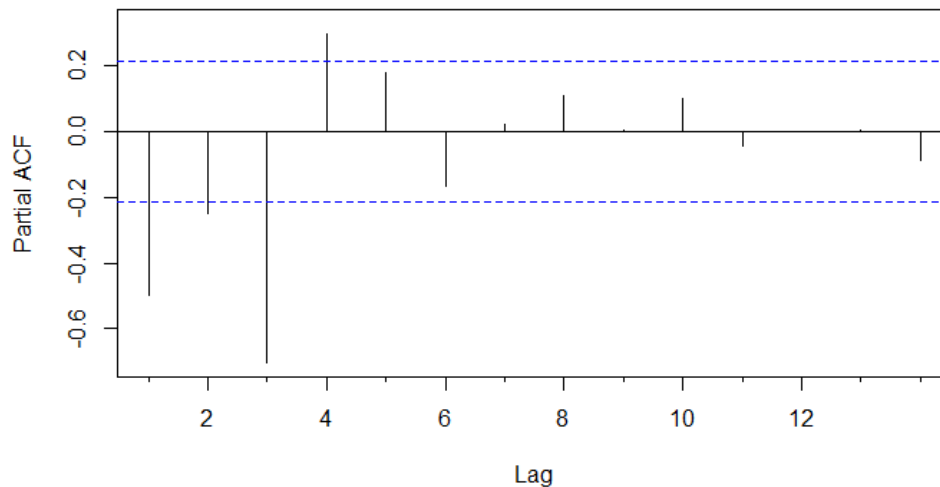


Figure 4.8: Funcion de Autocorrelacion Parcial.

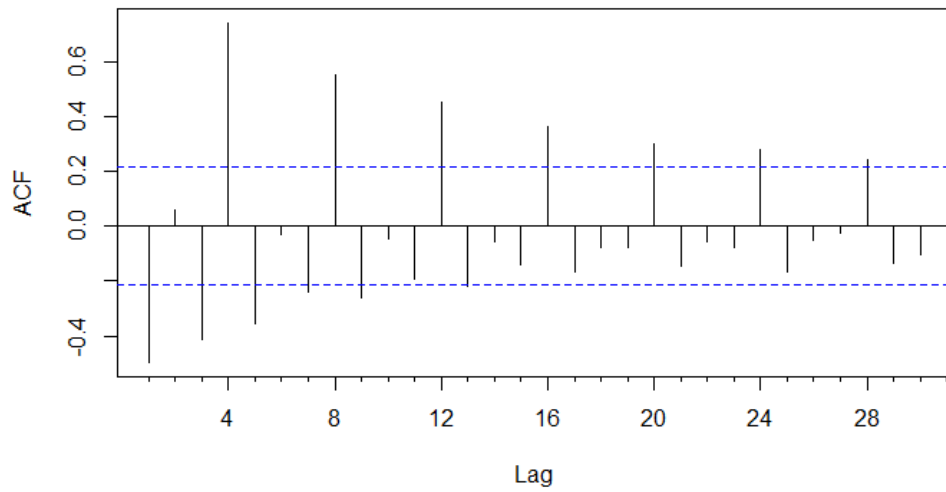


Figure 4.9: Funcion de Autocorrelacion.

En base a estas representaciones, decidimos utilizar un $p = 5$ y $q = 16$ para realizar nuestro modelo personalizado de ARIMA. Para asegurarnos de que estos valores fueran los mas óptimos, decidimos implementar los modelos AR y MA, y así poder asegurar los ordenes a utilizar. Los gráficos de los resultados de estas funciones se pueden ver en la Figura 4.10 y Figura 4.11, respectivamente.

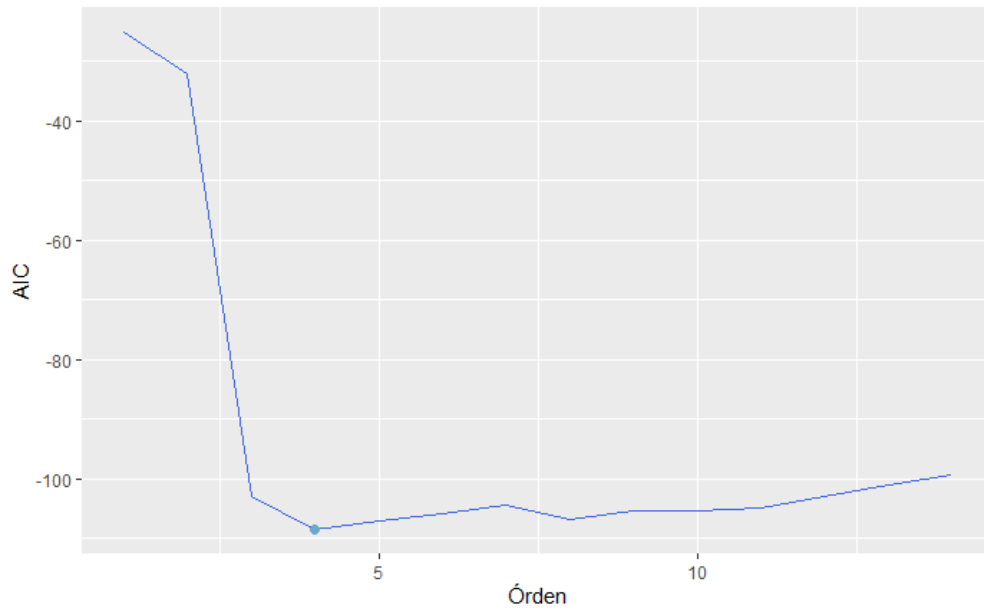


Figure 4.10: Gráfico AR.

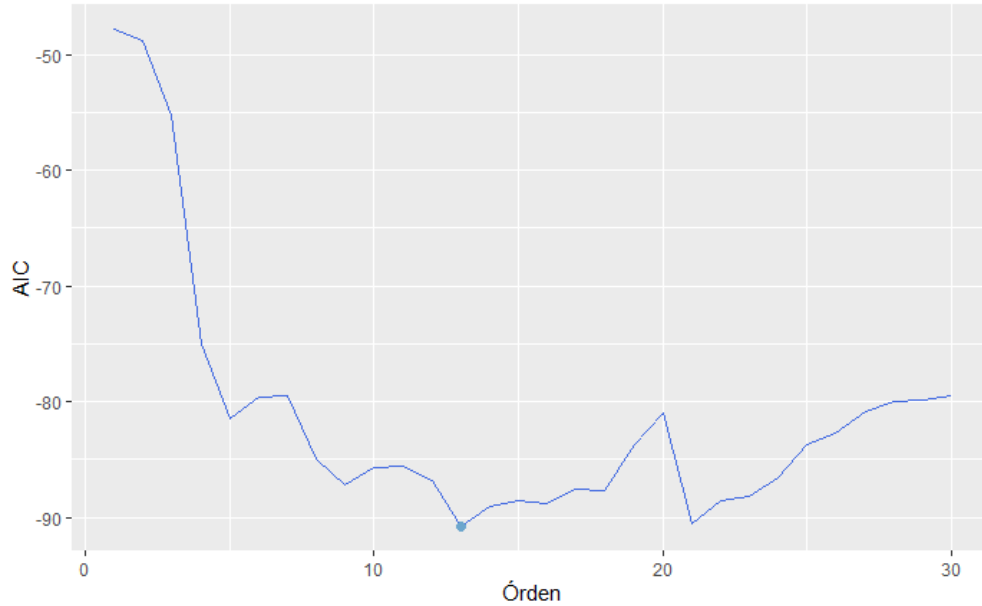


Figure 4.11: Gráfico MA.

Estos modelos nos dieron resultados diferentes a los que habíamos determinado previamente, siendo el valor mínimo para $p = 4$ y para $q = 13$. A pesar de los valores obtenidos, decidimos continuar también con nuestros valores determinados anteriormente, por lo cual realizamos dos modelos ARIMA personalizados, para así poder comparar los resultados obtenidos contra el modelo automático.

4.5 Quinto punto

Consigna: ¿Cuáles son los parámetros obtenidos para el modelo ARIMA automático?

Al implementar el modelo ARIMA automático, obtuvimos la siguiente salida: $\text{ARIMA}(0,0,1)(0,1,1)[4]$. En base a esto, podemos determinar que el ARIMA automático encontró dos conjuntos de parámetros con los que pudo obtener los mejores resultados. Para el primer conjunto tiene $p=0$, $d=0$ y $q=1$, mientras que para el segundo tiene $p=0$, $d=1$ y $q=1$.

4.6 Sexto punto

Consigna: Predecir las ganancias del último año utilizando los dos modelos ARIMA hallados. Calcular el criterio de información de Akaike (AIC) y el error de porcentaje medio absoluto (MAPE) en cada caso y decidir, en función de estos valores, qué modelo realiza las mejores predicciones.

Realizando la implementación de los dos modelos personalizados y el modelo automático, procedimos a predecir los valores de ganancia para el último año (1980) mediante el conjunto de prueba. Luego de predecir dichos valores con cada modelo, realizamos el cálculo de AIC y MAPE, ambos resultados se encuentran detallados en la Tabla 11.

Modelo	AIC	RMSE	MAPE
ARIMA(13,0,4)	-99.11	0.184	102.6
ARIMA(16,0,5)	-98.76	0.144	80.04
ARIMA Automático	-113.36	0.172	95.7

Table 11: Resultados de errores de los modelos.

Teniendo en cuenta los resultados obtenidos para cada modelo, vemos que el modelo de ARIMA personalizado con los parámetros determinados por las funciones de AR Y MA, nos generó el peor valor de RMSE Y MAPE, siendo valores mucho mas grandes que los otros dos modelos. A pesar de esto, el valor de AIC que obtuvo no es significativamente malo en comparación con el automático, quien obtuvo el peor rendimiento en esa medida, mientras que para las otras dos obtuvo un resultado mas bajo de error en comparación al primero. Por ultimo, el que mejor resultados obtuvo para las 3 métricas fue el modelo ARIMA personalizado con los parámetros que visualizamos en los gráficos de PACF y ACF. Por lo tanto, podemos concluir que el modelo que mejor se ajusta a los datos es el ARIMA(13,0,4).