

**Universidad Nacional De Colombia**  
**Sede Medellín**



**Facultad de Ciencias**  
Departamento de Estadística

**Trabajo 2**  
Identificación, Estimación y Validación de Modelos  
ARIMA

**Juan David García Zapata**  
**Juan Camilo Miranda Paz**

Series de tiempo Univariadas

**Agosto, 2024**

# Índice

<b>1. Punto 1</b>	<b>2</b>
1.1. Literal A) . . . . .	4
1.2. Literal B) . . . . .	6
1.3. Literal C) . . . . .	8
1.4. Literal D) . . . . .	13
1.5. Literal E) . . . . .	14
1.6. Literal F) . . . . .	15
1.7. Literal G) . . . . .	20
<b>2. Punto 2</b>	<b>21</b>
2.1. Literal A) . . . . .	21
2.2. Literal B) . . . . .	36

# 1. Punto 1

## Enunciado

Considere la suma del último dígito de la cédula de cada uno de los integrantes del grupo y divida esta suma por 4 y considere el conjunto de datos que corresponde con el residuo de dicha división, seleccione la base de datos acorde:

- A) Verifique si es necesario transformar los datos con una transformación Box-Cox.
- B) Realice Pruebas de Raíces unitarias.
- C) Realice usted mismo la identificación de un modelo ARIMA para la serie, teniendo en cuenta los autocorrelogramas ACF, PACF y función de autocorrelación muestral extendida ESACF.
- D) Realice la identificación de **otro modelo** tentativo, esta vez utilizando la rutina de selección automática `auto.arima()` de la librería `forecast` **En los siguientes ítems, considere el modelo que usted mismo estimó.**
- E) Lleve a cabo una estimación de los parámetros del modelo.
- F) Valide o realice un diagnóstico de los modelos que ha logrado identificar en el anterior ítem.
- G) Finalmente, describa la tendencia de la serie analizada y responda:
  - 1) La tendencia de la serie posee componentes determinística y aleatoria, solo una de ellas (cual) o ambas?
  - 2) El proceso adecuado para modelar la serie se trata de un proceso estacionario en tendencia o un proceso de diferencias estacionarias?

## Librerías Utilizadas

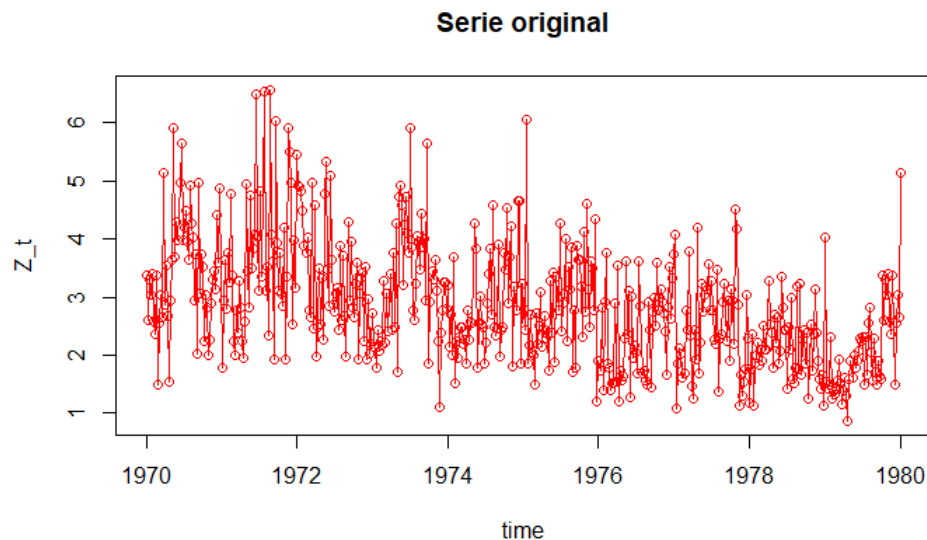
```
library(ggplot2)
library(tseries)
library(car)
library(TSA)
library(foreign)
library(lmtest)
library(quantmod)
library(astsa)
library(urca)
library(fANCOVA)
library(fUnitRoots)
library(forecast)
```

En este punto se analizarán los datos correspondientes a la base de datos **Serie02\_so\_SS** dado que la último dígito de la cédula de cada integrante es 0 (En el caso de Juan David) y 2 (En el caso de Juan Camilo) por lo tanto el residuo que se obtiene al dividir la suma entre 4 es  $2(2 \% 4 = 2)$

```
z <- read.csv("C:\\Users\\Juan\\Documents\\Serie02_so_SS - Serie02_so_SS.csv")
z1 <- ts(z,start = 1970, end = 1980, frequency = 52)
```

Antes de empezar con los literales, observamos la gráfica de la serie que abarca las fechas que comienzan en el año 1970 hasta 1980, con 508 observaciones.

```
plot.ts(z1, ylab = "Z_t",main = " Serie original ", xlab = "time", type = "o",
col= 'red', lwd = 1)
```



De acuerdo con la visualización de la serie temporal, se pueden hacer las siguientes observaciones:

### 1. No estacionariedad en la media y la varianza

**Media:** Aunque no es evidente una tendencia extremadamente clara, hay una ligera tendencia descendente en la serie temporal. Esto sugiere que la media de la serie no es constante a lo largo del tiempo, lo cual es un signo de no estacionariedad.

**Varianza:** La varianza también parece no ser constante. En ciertos periodos (por ejemplo, alrededor de 1970-1974), las fluctuaciones parecen más pronunciadas en comparación con otros periodos, lo que sugiere una varianza heterogénea.

## 2. Tendencia

La tendencia general de la serie parece ser ligeramente negativa, aunque no es altamente pronunciada. Este tipo de tendencia podría indicar que el proceso subyacente tiene algún tipo de disminución en su nivel medio a lo largo del tiempo.

## 3. Ciclicidad y fluctuaciones

Se observan fluctuaciones recurrentes a lo largo del tiempo, pero no hay una evidencia clara de patrones cíclicos definidos. Esto podría sugerir la presencia de componentes estocásticos o aleatorios en la serie, sin ciclos bien definidos.

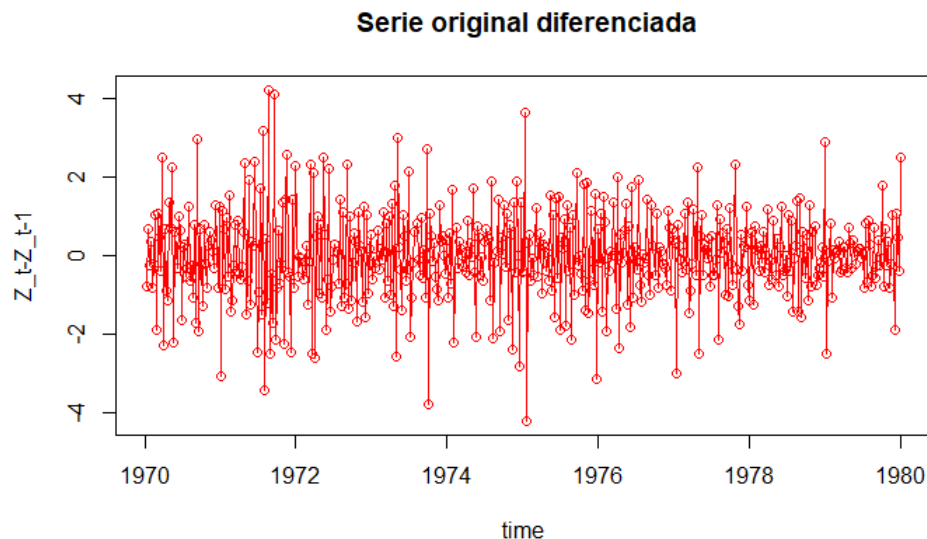
## 4. Picos y caídas abruptas

En ciertos puntos, la serie presenta picos y caídas abruptas, lo que podría ser indicativo de eventos específicos o choques en el sistema subyacente.

### 1.1. Literal A)

Analizaremos la varianza de los datos tras aplicar diferenciación para estabilizar la serie temporal, eliminando tendencias y estacionalidades y facilitando un análisis más claro de la variabilidad.

```
plot.ts(diff(z1),main = " Serie original diferenciada",ylab = "Z_t-Z_t-1",  
xlab = "time", type = "o",col= 'red')
```



De acuerdo con la gráfica de la serie diferenciada, la varianza no parece ser constante, ya que se observan tramos con mayor variabilidad en comparación con otros. Esto sugiere la necesidad de aplicar una transformación adicional, como la transformación de Box-Cox, para estabilizar la varianza y mejorar la interpretabilidad y el análisis de la serie temporal.

## Transformación de Box-Cox

```
# Primer metodo
(tBoxCox=powerTransform(z1))
summary(tBoxCox)
# Segundo metodo
BoxCox.lambda(z1, method=c("loglik"), lower=-2, upper=2)
# Tercer metodo
BoxCox.lambda(z1, method=c("guerrero"))
```

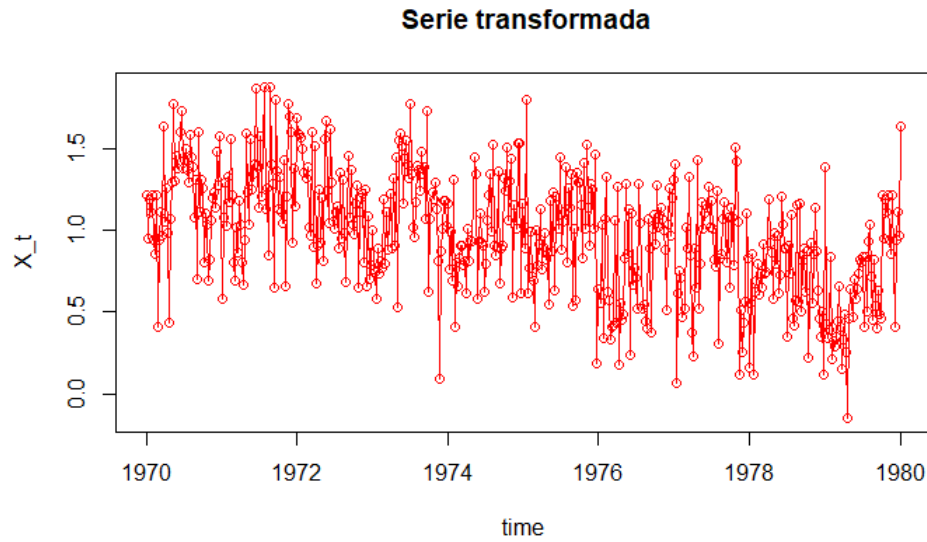
Método	$\lambda$ estimada
powerTransform(z1)	0.1414186
BoxCox.lambda(z1, method=c("loglik"))	0.1
BoxCox.lambda(z1, method=c("guerrero"))	0.1632403

Tabla 1: Estimación del parámetro de transformación ( $\lambda$ ) con diferentes métodos.

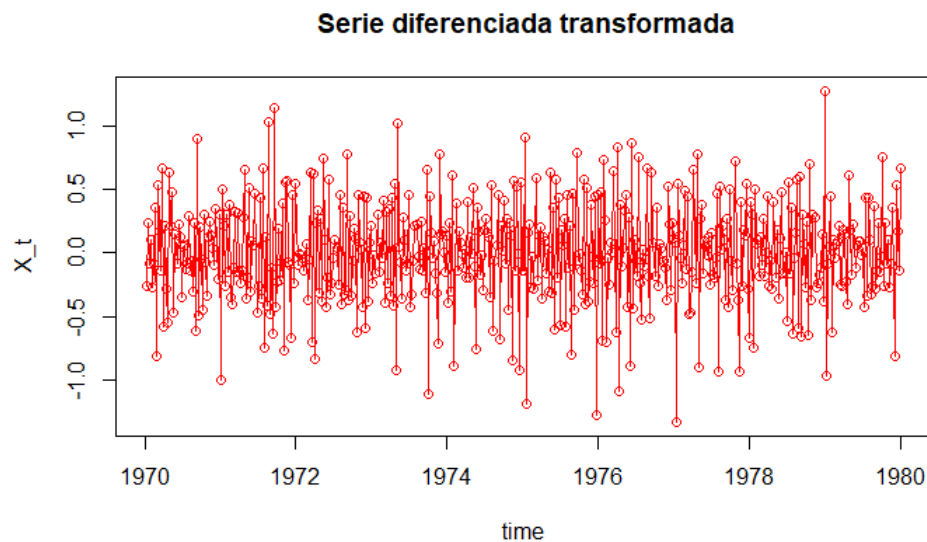
Potencia	Transformación	Descripción
$\lambda_1 = 2$	$Y' = Y^2$	Cuadrado
$\lambda_1 = 1$	$Y' = Y$	Datos sin Transformar
$\lambda_1 = 0.5$	$Y' = \sqrt{Y}$	Raíz Cuadrada
$\lambda_1 = 0.333$	$Y' = \sqrt[3]{Y}$	Raíz Cúbica
$\lambda_1 = 0$	$Y' = \ln(Y)$	Logaritmo
$\lambda_1 = -0.5$	$Y' = \frac{1}{\sqrt{Y}}$	Raíz Cuadrada Inversa
$\lambda_1 = -1$	$Y' = \frac{1}{Y}$	Reciproco

De acuerdo con los valores de  $\lambda$  obtenidos de los tres métodos, todos convergen, al ser redondeados, a un valor cercano a 0.1. Basándonos en esta observación y considerando que  $\lambda$  es aproximadamente 0, se considera aplicar una transformación logarítmica a los datos, ya que es la más adecuada en este caso para estabilizar la varianza.

```
#Los datos Transformados
Z=log(z1)
plot.ts(Z ,main = " Serie transformada",ylab = "X_t", xlab = "time", type = "o",
col='red', lwd = 1)
```



```
plot.ts(diff(Z) ,main = " Serie diferenciada transformada",ylab = "X_t",
xlab = "time",
type = "o",col='red', lwd = 1)
```



Después de aplicar la transformación logarítmica a los datos de la serie, se observa una mejora en la varianza, lo que confirma la necesidad de transformar los datos. A partir de ahora, todos los procedimientos posteriores se realizarán utilizando los datos transformados.

## 1.2. Literal B)

Para realizar la prueba de raíces unitarias se realizara la prueba de Dickey-Fuller, en la que el juego de hipótesis esta dada por:

$$H_0 : \phi = 1$$

$$H_i : \phi < 1$$

Estadístico de prueba:

$$T = \frac{\hat{r}}{\hat{S}_e(\hat{r})}$$

Región de rechazo:

$$T < T_{critico}(\alpha)$$

En esta prueba  $H_0$  establece que la serie tiene una Raíz unitaria, lo que quiere decir que la serie no es estacionaria, por lo tanto rechazar  $H_0$  indicaría que la serie es estacionaria. Para la prueba de Dickey- Fuller vamos a considerar en primera instancia el siguiente modelo:

$$\Delta Z_t = \beta_0 + \beta_1 t + \gamma Z_{t-1} + at$$

Con la ayuda de R tenemos los siguiente coeficientes:

```
(maxlag=floor(12*(length(Z)/100)^(1/4)))
[1] 18
ru_tz=ur.df(Z, type = c("trend"), lags=maxlag, selectlags = c("BIC"))
summary(ru_tz)
```

Coeficiente	Estimación	Error Estándar	Valor t	Pr(< t )
(Intercept)	0.8177864	0.0804547	10.165	< 2e-16***
z.lag.1	-0.6173835	0.0562734	-10.971	< 2e-16***
tt	-0.0007941	0.0001203	-6.599	1.06e-10***
z.diff.lag	-0.2402843	0.0439290	-5.470	7.14e-08***

Tabla 2: Resumen de los coeficientes del modelo.

El modelo indica que tanto el coeficiente del intercepto como el de la **tendencia** son significativos. Sin embargo, nuestro principal interés recae en el valor de  $\gamma$ , representado por el coeficiente de **z-lag.1**. Es importante recordar que no podemos utilizar el valor  $p$  como criterio decisivo, ya que el estadístico  $t$  no sigue una distribución convencional en este contexto. En su lugar, estamos tratando con una distribución especial donde los valores críticos dependen del tamaño de la muestra, por lo que debemos enfocarnos en estos valores críticos para nuestra interpretación.

Tenemos entonces que  $T = -10.9711$ , ahora, los **valores críticos**:

Nivel de Significancia	1 %	5 %	10 %
$\tau_3$	-3.96	-3.41	-3.12



**Region de rechazo** en cada nivel de significancia:

	1 %	5 %	10 %
$T < T_{critico}(\alpha)$	-10.9711 < -3.96	-10.9711 < -3.41	-10.9711 < -3.12
$T < T_{critico}(\alpha)$	Se Rechaza	Se Rechaza	Se Rechaza

En todos los niveles de significancia considerados (1 %, 5 %, y 10 %), se rechaza la hipótesis nula, lo que nos lleva a concluir que la serie no sigue un proceso estocástico, es decir, no presenta raíces unitarias. Esto implica que la serie es estacionaria. La ausencia de raíces unitarias indica que los valores de la serie tienden a retornar a su media a lo largo del tiempo, haciendo que su comportamiento sea más predecible y adecuado para el análisis y modelado estadístico.

**Observación:** Este análisis anterior asume que la serie es estacionaria, sin embargo, al observar la gráfica de la serie original, la estacionariedad no es clara. Además, la media no parece regresar a un valor constante. Si consideramos el resultado de la prueba de Dickey-Fuller y los coeficientes en el Tabla 2, donde la tendencia es significativa, se resalta una contradicción. por lo que esto puede conllevar a un modelo un poco diferente al que se tenía en mente desde un principio

### 1.3. Literal C)

Para la identificación del modelo, es importante considerar los aspectos descritos anteriormente. Dado que no se observan raíces unitarias y la tendencia es negativa, esto sugiere que la tendencia es determinística, no estocástica. Por lo tanto, el modelo propuesto es un Modelo de regresión lineal con errores estructurales ARMA:

$$Z_t = \beta_0 + \beta_1 t + E_t$$

Donde:

1.  $\beta_0 + \beta_1 t$  representa la parte determinística del modelo, que captura la tendencia lineal en los datos.
2.  $E_t$  por definir el orden de p y q para modelar un ARMA.

Para modelar las características adicionales de los residuos  $E_t$  se puede aplicar un modelo ARIMA, que permitirá capturar las dependencias temporales en los errores residuales.

Aplicaremos una regresión lineal para modelar la relación entre la serie temporal y el tiempo.

```
t<- 1:length(Z)
#variables para el ajuste, en el polinomio de grado 1
poli=Mipoly(tiempo=t,grado=1)
X1=data.frame(poli)
```

```

# Regresion simple
mod1 <- lm(Z~.,data = X1)
summary(mod1)

#Estimar los errores
yhatmod1=ts(exp(fitted(mod1))*exp(summary(mod1)$sigma^2/2),freq=52,start=start(Z))
res2 <- residuals(mod1)

```

Coefficients	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.291	2.830e-02	45.63	<2e-16***
<i>t</i>	-1.192e-03	9.395e-05	-12.69	<2e-16***

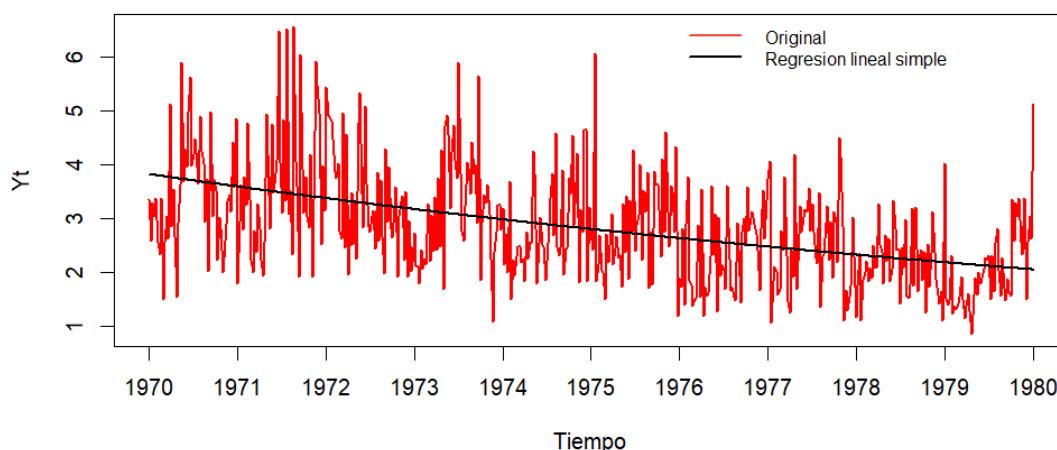
Tabla 3: Summary of the regression coefficients

Después de aplicar el modelo, se generó la gráfica comparativa para visualizar cómo el ajuste del modelo (en rojo) describe la serie temporal original (en negro). Esta visualización nos permite evaluar la calidad del ajuste y verificar si el modelo captura adecuadamente las características principales de la serie, como la tendencia.

```

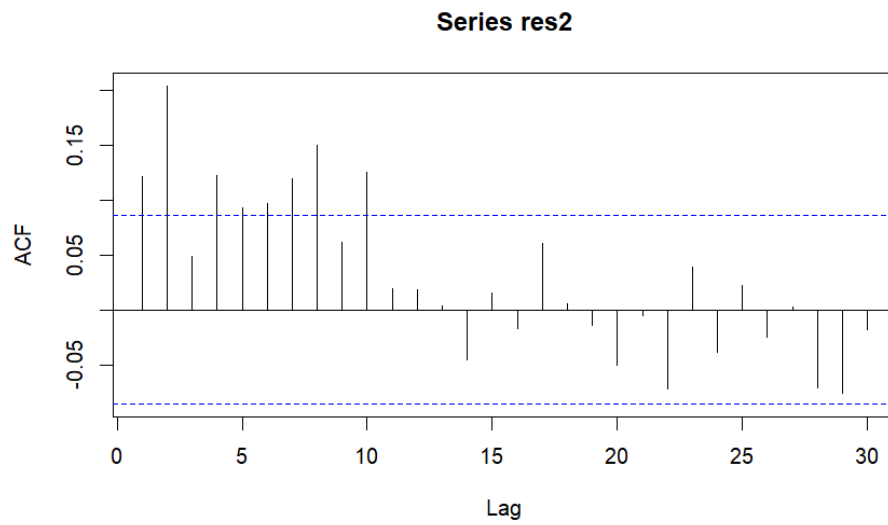
#Grafico de la serie y su ajuste final
plot(exp(Z), xlab="Tiempo", ylab="Yt", lwd=2, xaxp=c(1970,1980,10), col="red")
lines(yhatmod1,col="black",lwd=2)
legend("topright",legend=c("Original","Regresion lineal simple"),col=c(2,1),
lty=1, bty = "n",cex=0.8,lwd=2)

```



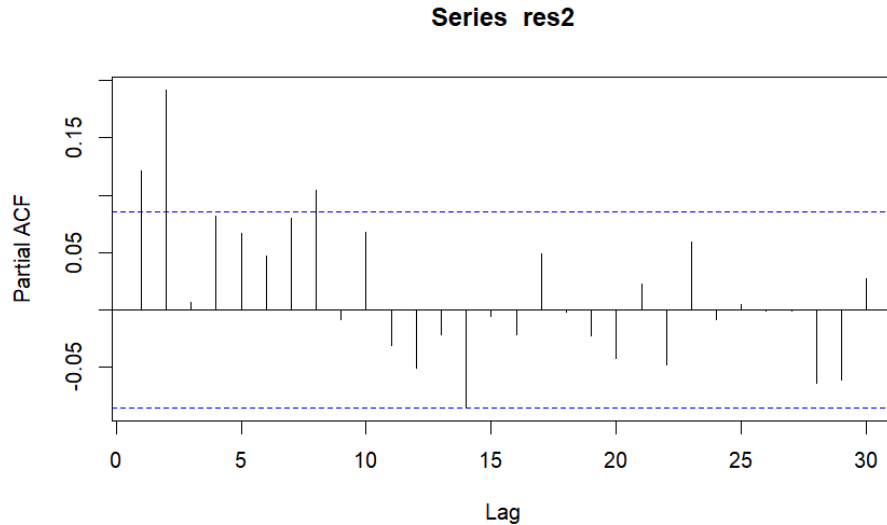
Ahora nuestro enfoque esta en los residuos del modelo resultante ya que este nos permitira identificar el ARIMA, en este caso un posible ARMA o AR o MA ya que el ARIMA como tal esta descartado por Dickey-Fuller.

```
acf(res2, lag.max=30)
```



**ACF:** de los residuales

```
pcf(res2, lag.max=30)
```



**PACF:** de los residuales

Los gráficos de ACF y PACF proporcionan información sobre la estructura temporal de los residuos:

**ACF (Función de Autocorrelación):** El gráfico de la ACF muestra un patrón de decaimiento y oscilante, lo que es característico de un proceso autorregresivo (AR). El decaimiento en cola sugiere que los residuos no son ruido blanco, sino que tienen una estructura temporal que podría ser modelada por un proceso AR.

**PACF (Función de Autocorrelación Parcial):** En el gráfico de la PACF, se observa un corte claro en el rezago 2, con un valor significativo en este punto. Esto indica que un modelo AR de orden 2 (AR(2)) podría ser adecuado para capturar la dependencia temporal en los residuos. Aunque se observa un valor significativo en un rezago más alto, es probable que estos sean el resultado de errores de estimación o ruido, y no indicativos de una estructura más compleja.

`eacf(res2, 18,18)`

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	x	x	x	x	x	x	x	x	x	o	x	o	o	o	o	o	o	o	o
1	x	x	x	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o	o
2	x	x	x	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o	o
3	x	x	x	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o	o
4	x	x	x	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o	o
5	x	x	x	x	o	o	o	o	o	x	o	o	o	o	o	o	o	o	o
6	x	o	x	x	o	x	o	o	o	x	o	o	o	o	o	o	o	o	o
7	x	o	x	x	x	x	o	o	o	x	o	o	o	o	o	o	o	o	o
8	x	x	x	x	x	o	x	o	o	x	o	o	o	o	o	o	o	o	o
9	x	x	x	x	x	o	x	o	x	x	o	o	o	o	o	o	o	o	o
10	x	x	x	x	x	x	x	x	x	x	o	o	o	o	o	o	o	o	o
11	x	x	x	x	x	x	x	x	x	x	o	o	o	o	o	o	o	o	o
12	x	x	x	x	x	x	o	x	o	x	x	o	o	o	o	o	o	o	o
13	x	x	x	o	o	x	o	x	x	o	o	x	o	o	o	o	o	o	o
14	x	x	x	o	o	o	o	x	x	o	o	x	x	x	o	o	o	o	o
15	x	x	x	o	x	o	o	x	o	o	x	x	x	x	o	o	o	o	o
16	x	o	x	x	o	x	o	o	x	o	x	o	x	x	o	o	o	o	o
17	x	x	o	x	o	x	x	o	o	o	o	o	x	x	x	o	o	o	o
18	x	o	o	x	x	x	x	x	o	o	o	o	x	o	x	o	o	o	o

**EACF** de los residuales

Con base en la matriz EACF y asumiendo que se busca un modelo que minimice los órdenes de AR y MA, se pueden construir varios modelos ARMA (p, q) que podrían ser adecuados. Sin embargo, los resultados de los gráficos ACF y PACF muestran un modelo claro, el cual es un ARMA (2,0). En términos del modelo ARIMA, esto se expresaría como un ARIMA (2,0,0).

Es importante destacar que este modelo AR (2,0) se aplicará a los residuos del modelo de regresión lineal con errores estructurales ARMA. La estructura temporal observada en los residuos sugiere que un componente autorregresivo de orden 2 es suficiente para capturar la dependencia temporal, lo que permite un modelado más preciso y robusto de los residuos del modelo de regresión.

**Modelo propuesto para los residuales: ARMA (2,0)**

$$Z_t = \beta_0 + \beta_1 t + E_t$$

Donde:

1.  $\beta_0 + \beta_1 t$  representa la parte determinística del modelo, que captura la tendencia lineal en los datos.
2.  $E_t = \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + a_t$  ó  $E_t = (1 - \phi_1 B + \phi_2 B^2) \epsilon_t + s_t$

Con  $a_t$  como los errores del modelo ARMA (2,0)

## 1.4. Literal D)

En este caso, consideraremos dos enfoques: primero, aplicaremos la función ‘auto.arima’ a la serie original transformada; segundo, utilizaremos ‘auto.arima’ sobre el modelo de regresión con errores estructurales ARMA.

*#Primer enfoque*

```
auto.arima(Z,max.p=5,max.q=5)
```

Series: Z

```
ARIMA(2,1,1)
```

Coefficients:

	ar1	ar2	ma1
	0.0363	0.1412	-0.8846
s.e.	0.0642	0.0599	0.0435

sigma^2 = 0.09568: log likelihood = -126.82

AIC=261.64 AICc=261.72 BIC=278.66

El modelo ARIMA(2,1,1) ajustado incluye un término de diferenciación, lo que contrasta con el modelo previamente considerado, donde no se incluyó un factor de diferenciación debido a que fue rechazado en la prueba de Dickey-Fuller. Este rechazo indicaba que la serie no requería diferenciación, lo que motivó la consideración de un modelo alternativo sin el componente de diferenciación. A pesar de esto, el ARIMA(2,1,1) muestra coeficientes significativos ( $ar1 = 0.0363$ ,  $ar2 = 0.1412$ ,  $ma1 = -0.8846$ ) y un buen ajuste con un AIC de 261.64 y un BIC de 278.66, lo que sugiere que este modelo también podría capturar adecuadamente la estructura temporal de la serie, aunque con un enfoque distinto. Modelos alternativos que ofrezcan un mejor ajuste.

*#Segundo enfoque*

Series: res2

```
ARIMA(2,0,2) with zero mean
```

Coefficients:

	ar1	ar2	ma1	ma2
	0.1013	0.6817	0.0083	-0.4843
s.e.	0.1563	0.1453	0.1731	0.1466

sigma^2 = 0.09256: log likelihood = -117.43

AIC=244.86 AICc=244.97 BIC=266.14

El modelo ARIMA(2,0,2) ajustado a los residuales del modelo Modelo de regresión lineal con errores estructurales sin diferenciación y con media cero muestra coeficientes significativos para los términos autorregresivos ( $ar1 = 0.1013$ ,  $ar2 = 0.6817$ ) y de media móvil ( $ma1 = 0.0083$ ,  $ma2 = -0.4843$ ). La varianza del error ( $\sigma^2$ ) es 0.09256, lo que indica un ajuste preciso del modelo. El logaritmo de la verosimilitud es -117.43, con un AIC de 244.86, AICc de 244.97 y un BIC de 266.14. Estos valores sugieren que el modelo ARIMA(2,0,2) proporciona un buen ajuste a la serie t

## 1.5. Literal E)

Para estimación de los parametros del modelo recordemos que el modelo ARMA se aplicara a los residuos( $E_t$ ) del modelo de regresión lineal que presenta errores estructurales es decir:

$$E_t = (1 - \phi_1 B + \phi_2 B^2)\epsilon_t + a_t$$

Procedemos a realizar una estimación de suma condicional de cuadrados con ayuda de la función **Arima()**, en nuestro caso ARIMA (2,0,0).

```
modelo1 <- Arima(z1, order=c(2,0,0), xreg=as.matrix(X1), method=c("CSS-ML"))
summary(modelo1)
```

En este caso, hemos decidido incluir el parámetro ‘xreg’ en el modelo porque nos permite calcular los coeficientes de la parte determinística de la regresión, y luego aplicar un modelo ARMA(2,0) a los residuales del modelo de regresión simple. Esto se basa en la técnica de regresión con errores ARIMA, donde se ajusta primero un modelo de regresión a los datos y luego se modelan los residuales mediante un proceso ARIMA. Esto es útil para capturar tanto la estructura determinística como la estocástica en los datos.

El enfoque tiene la ventaja de poder explicar parte de la variabilidad de la serie mediante la regresión, lo que puede llevar a intervalos de predicción más ajustados y precisos. Además, al modelar los residuales con un proceso ARMA, se asegura que cualquier estructura autocorrelada en los errores sea adecuadamente capturada, mejorando la calidad general del modelo

Series: Z

Regression with ARIMA(2,0,0) errors

Coefficients:

	ar1	ar2	intercept	t
	0.1502	0.2446	1.2869	-0.0012
s.e.	0.0428	0.0428	0.0442	0.0002

sigma^2 = 0.09449: log likelihood = -122.75

AIC=255.5 AICc=255.62 BIC=276.78

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	0.0002386549	0.3062077	0.2447969	-20.56352	41.43237	0.6781264
ACF1	-0.007383541					

$$Z_t = 1.2869 - 0.0012t + E_t$$

con  $E_t$

$$E_t = (1 - 0.1502B - 0.2446B^2)\epsilon_t + a_t$$

o de forma expandida:

$$E_t = -0.1502\epsilon_{t-1} - 0.2446\epsilon_{t-2} + a_t$$

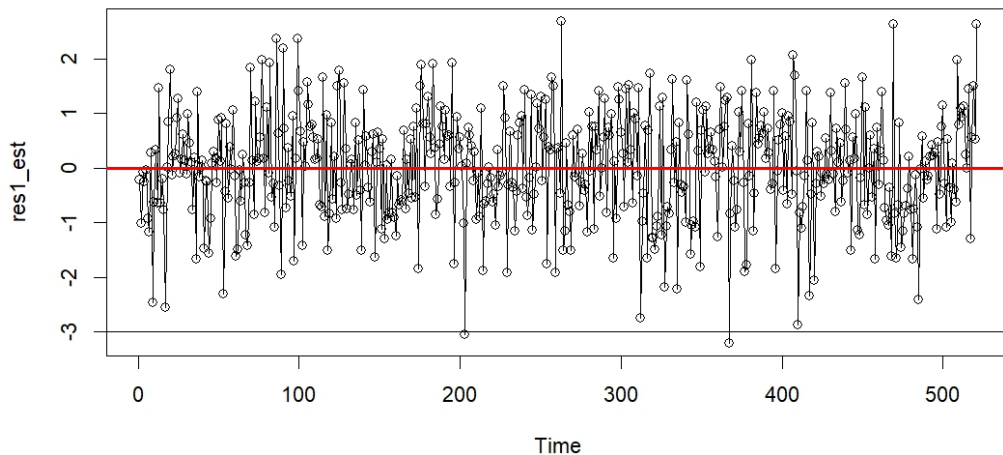
Donde  $a_t$  corresponden a los residuos del modelo ARMA(2,0) propuesto para los errores del modelo de regresión lineal con errores estructurales ARMA, además  $a_t$  debe cumplir unos supuestos para ser un ruido blanco.

## 1.6. Literal F)

El objetivo es verificar que los  $a_t$  se comportan como un ruido blanco gaussiano, donde validamos:

1. Media cero
2. Varianza constante
3. Ningún patrón de dependencia aparente

```
res1_CSS_ML_ML=residuals(modelo1)
res1_est=res1_CSS_ML_ML/(modelo1$sigma2^.5)
plot.ts(res1_est, type="o")
abline(a=0,b=0,lwd=3, col="red")
```



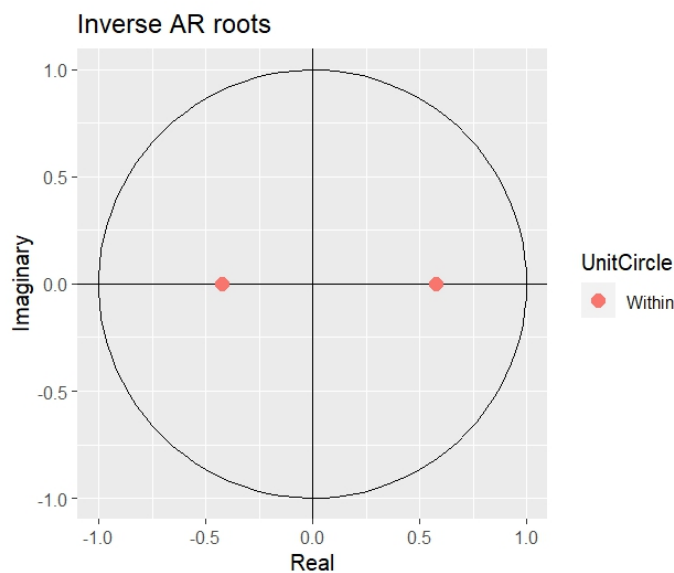
Gráfica de los residuos contra el tiempo



Observando la gráfica, podemos notar que aparentemente cumple con la condición de media cero y una posible varianza constante, sin embargo, hay tramos de la gráfica donde no es tan claro concluir esto. No se observan patrones de dependencia pero sí algunos datos atípicos.

También podemos ver que se comporta de forma estacional, para validar esta afirmación, evaluamos las raíces del polinomio utilizando la función `autoplot` en R.

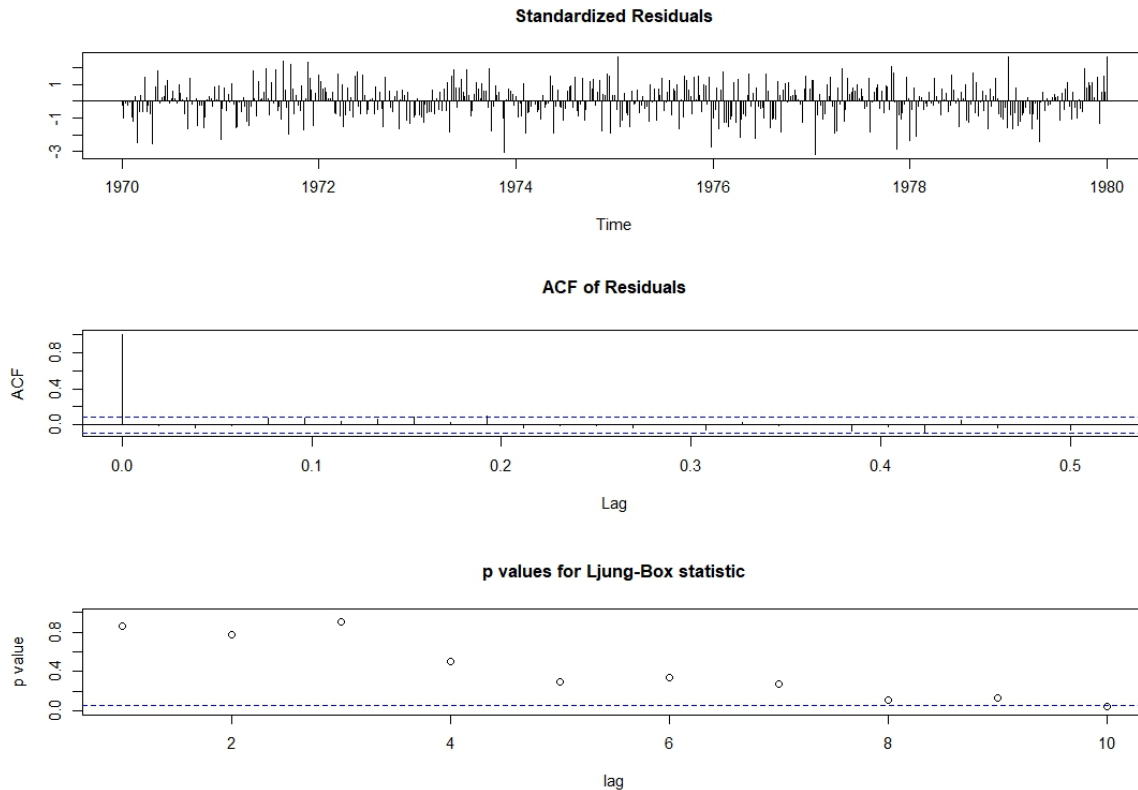
```
autoplot(modelo1)
```



Gráfica sobre las raíces del modelo

Con la gráfica concluimos que, dado que las raíces del modelo estarían por fuera del círculo unidad dado por la inversa, se comporta estacionario.

```
tsdiag(modelo1)
```



### Correlogramas y test

Observamos que los residuales estadarizados se asemejan a un comportamiento normal con algunos datos atípicos. Recordemos que para ser considerado un dato atípico, se consideran aquellos que pasen el umbral de  $(-3, 3)$ .

La ACF muestra que los residuales solo tienen correlación consigo mismos en el rezago 0, mientras que en los demás la correlación es aproximadamente cero. Esto indica que los residuos no presentan autocorrelación significativa, sugiriendo que el modelo ha capturado bien la estructura de los datos.

Por último, mirando la gráfica de Ljung-Box, notamos que en los primeros rezagos el valor  $P$  supera un  $\alpha$  de 0.05, lo que nos permite concluir que no hay autocorrelación en esos primeros, sin embargo, en rezagos más lejanos el valor  $P$  es inferior a este  $\alpha$ , pero al ser un rezago tan lejano, no influye en el dato siguiente.

### Datos atípicos

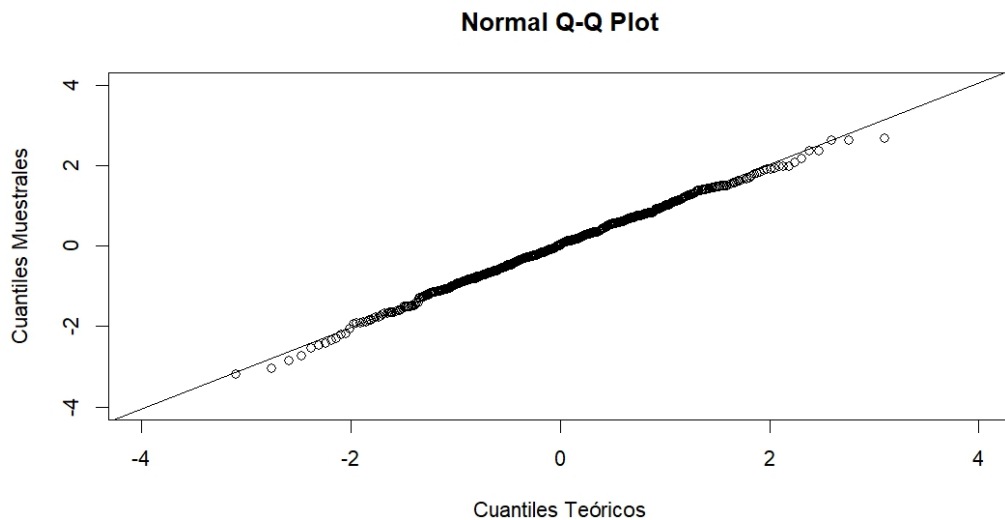
```
(Nobs_Esp=round(length(Z)*2*pnorm(-3, mean = 0, sd = 1, lower.tail = TRUE)))
[1] 1
```

```
ind=(abs(res1_est)>3.0)
sum(ind)
[1] 2
```

Aunque se detectaron dos observaciones atípicas que superan las 3 desviaciones estándar, lo que podría afectar la normalidad de los residuos, el número de datos atípicos es muy pequeño. Por lo tanto, es poco probable que estos pocos valores tengan un impacto significativo en el análisis global o en las suposiciones del modelo.

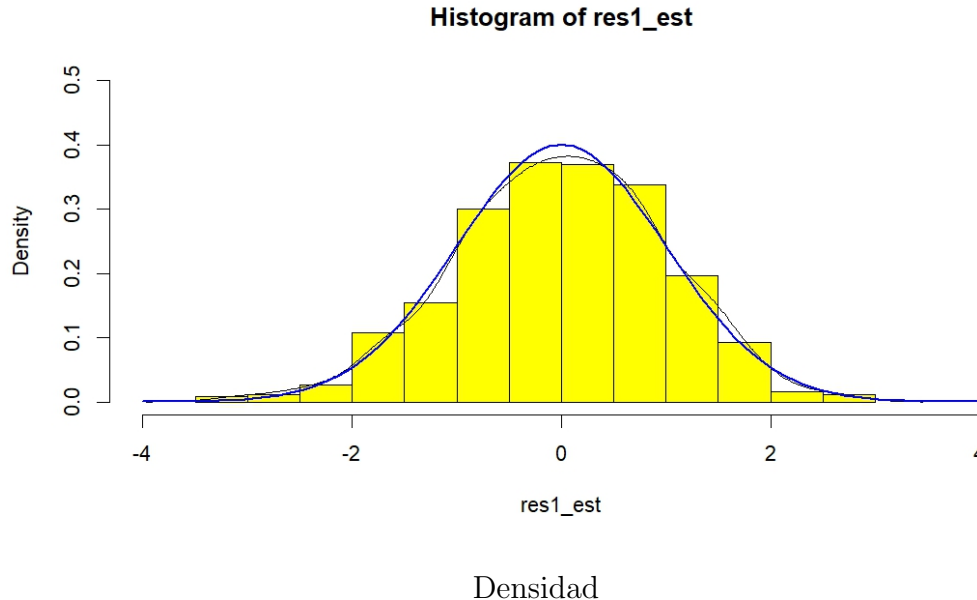
### Verificacion de normalidad

```
qqnorm(res1_est, xlab = "Cuantiles Teóricos", ylab = "Cuantiles Muestrales",
xlim=c(-4,4), ylim=c(-4,4))
qqline(res1_est)
```



Q-Q Plot

```
mu<-mean(res1_est)
sigm<-sd(res1_est)
x<-seq(-4,4,length=100)
y<-dnorm(x,mu,sigm)
hist(res1_est,prob=T,ylim=c(0,0.5),xlim=c(-4,4),col="yellow")
lines(density(res1_est))
lines(x,y,lwd=2,col="blue")
```



```
shapiro.test(res1_est)
```

Obtenemos la salida:

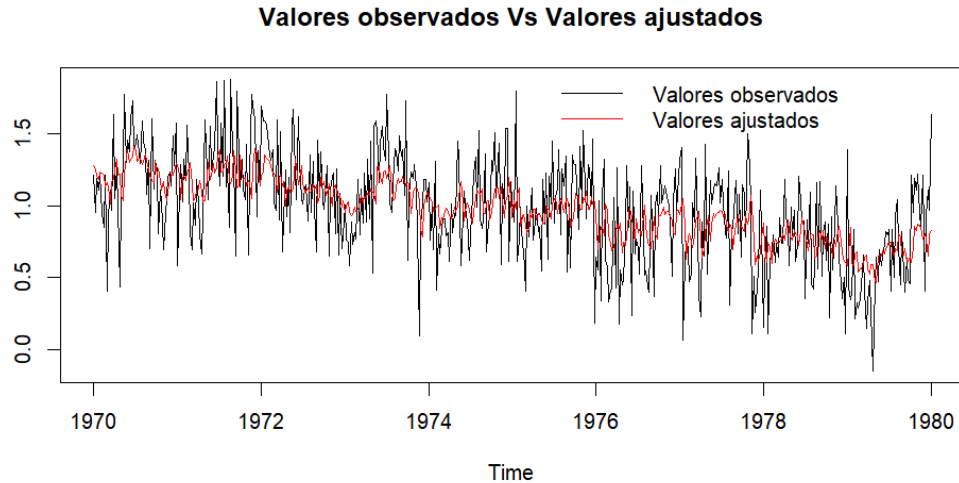
```
Shapiro-Wilk normality test
data:  res1_est
W = 0.9972, p-value = 0.5207
```

Finalmente, los residuales y la densidad de los residuales gráficamente se comportan como una distribución normal, y esto lo corroboramos con el test de Shapiro-Wilk. Después de realizar un análisis exhaustivo de los residuos generados por el modelo, hemos llegado a la conclusión de que los términos de error se comportan de acuerdo con una distribución normal. Esto sugiere que las suposiciones subyacentes del modelo, en cuanto a la normalidad de los errores, se mantienen y son válidas, lo que refuerza la fiabilidad y robustez de los resultados obtenidos.

### Valores observados Vs valores ajustados

```
res1_CSS_MLn <- residuals(modelo1)
ajustn <- Z - res1_CSS_MLn
ts.plot(Z, ajustn, col=c("black", "red"), lty=c(1,1),
main= "Valores observados Vs Valores ajustados")
lines(Z, col="black")
lines(ajustn, col="red")

legend("topright", legend=c("Valores observados", "Valores ajustados"),
col=c("black", "red"), lty=1, bty="n")
```



Densidad

El modelo ajustado logra captar de manera efectiva la tendencia general de la serie temporal, siguiendo de cerca los valores observados. Sin embargo, en ciertos periodos, se observan discrepancias donde el modelo no logra capturar completamente las fluctuaciones rápidas de los datos observados. Esto sugiere que, si bien el modelo es adecuado para describir la tendencia general, podría beneficiarse de refinamientos adicionales para mejorar su capacidad de captura de variaciones más rápidas o inusuales en la serie temporal.

### 1.7. Literal G)

- La tendencia de la serie posee componentes determinística y aleatoria, solo una de ellas (cúal) o ambas?

Dado los resultados obtenidos en los anteriores ítems, consideramos que el modelo incluye tanto una parte determinística como una parte aleatoria. La componente determinística podría estar capturando una tendencia o patrón predecible, mientras que la parte aleatoria o estocástica captura las fluctuaciones aleatorias alrededor de esa tendencia.

- El proceso adecuado para modelar la serie se trata de un proceso estacionario en tendencia o un proceso de diferencias estacionarias?

Como ajustamos un modelo que incluye tanto una parte determinística, es decir, la tendencia, y además una aleatoria, donde los residuos son estacionarios, concluimos que el proceso adecuado para modelar la serie sería un proceso estacionario en tendencia.

## 2. Punto 2

### Enunciado

Considere la suma del último dígito de la cédula de cada uno de los integrantes del grupo y divida esta suma por 4 y considere el conjunto de datos que corresponde con el residuo de dicha división, seleccione la base de datos acorde. Luego transformela a un objeto de la clase `time series` con la función `ts()` y proceda a desarrollar los ejercicios.

- a) Realice los plots que usted considere adecuados y necesarios.
- b) Identifique y ajuste un modelo ARIMA **estacional** adecuado, **teniendo cuidado de dejar las observaciones correspondientes al último año para la evaluación de pronósticos**. Utilice los autocorrelogramas para la identificación.

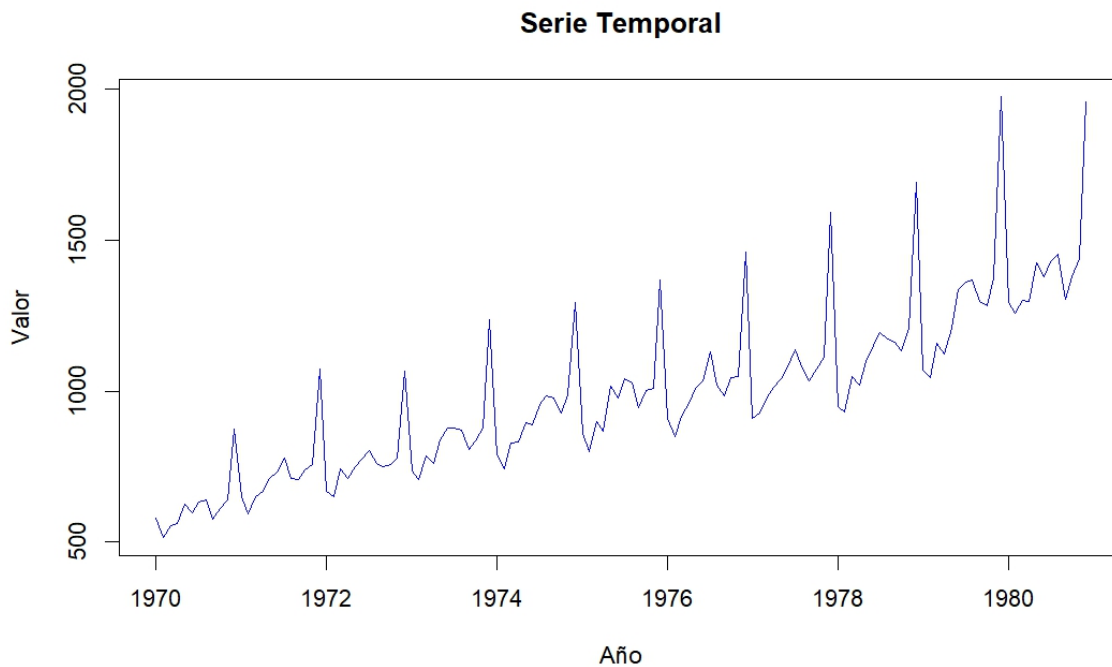
### 2.1. Literal A)

En este punto se analizarán los datos correspondientes a la base de datos **LicorVents-Weip184**. Dado que la último dígito de la cédula de cada integrante es 0 (En el caso de Juan David) y 2 (En el caso de Juan Camilo) por lo tanto el residuo que se obtiene al dividir la suma entre 4 es  $2(2 \% 4 = 2)$

Antes de empezar con los literales, observamos la gráfica de la serie que abarca las fechas que comienzan en el año 1970 hasta 1980, con 132 observaciones.

```
Z <- read_excel("Licor_Vents_Wei_p184.xlsx", col_names = FALSE)
Z = as.data.frame(Z)

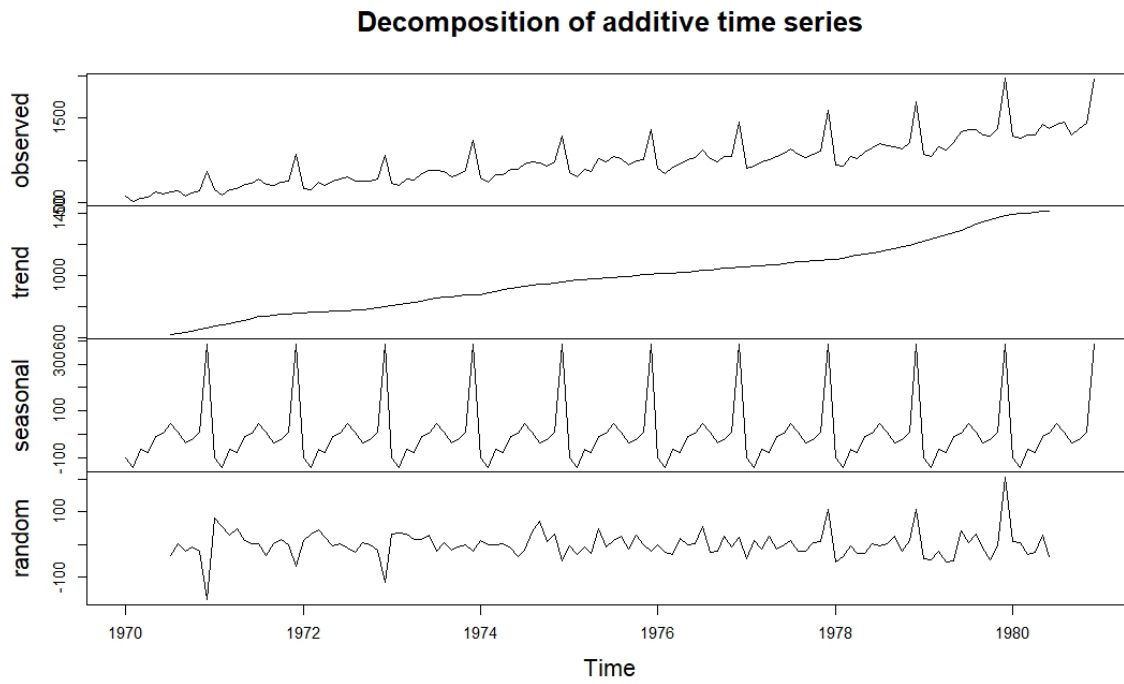
data_vector <- as.vector(t(Z))
ts_data <- ts(data_vector, start = c(1970, 1), frequency = 12)
plot(ts_data, main = "Serie Temporal", xlab = "Año", ylab = "Valor",
col = "blue")
```



De la gráfica podemos observar que no es estacionaria, tiene una tendencia positiva y su varianza no es constante y tiene forma multiplicativa. Además, presenta una periodicidad cada 12 estaciones, es decir, cada año.

Para confirmar lo anterior, vamos a verificar la estacionalidad y la tendencia descomponiendo la serie, así:

```
des.serie <- decompose(ts_data)
plot(des.serie)
```



Descomposición de la serie de tiempo en sus diferentes componentes

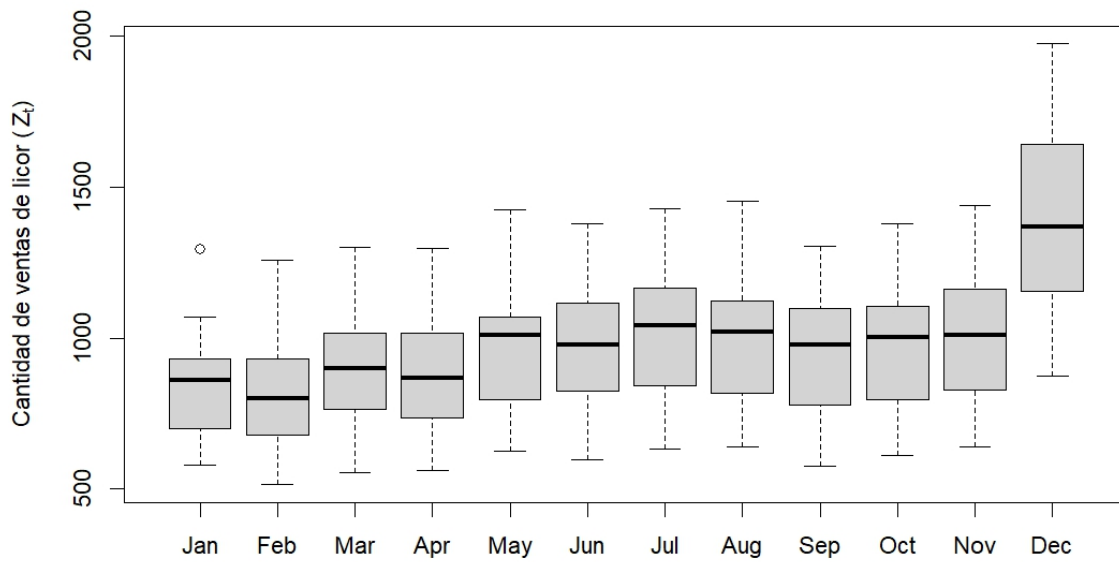


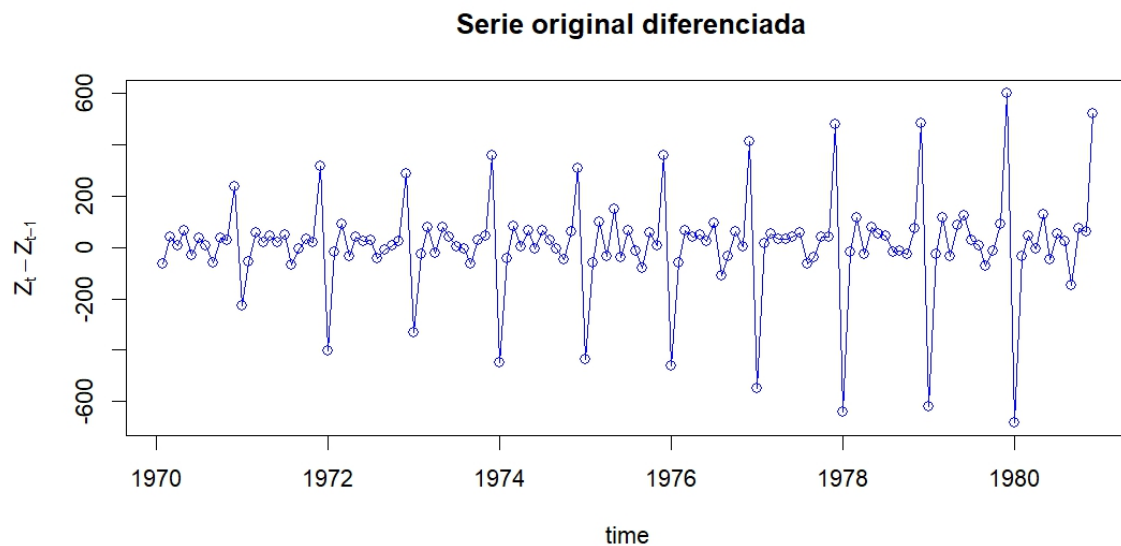
Diagrama de cajas por meses



Con la gráfica observada, podemos notar que no hay un crecimiento notable de la media a lo largo de los meses, sin embargo, hay una gran diferencia en el momento que pasa al mes de diciembre. Con esto podemos confirmar que hay una tendencia positiva y una estacionalidad relacionada con el mes de diciembre que es donde se disparan las ventas de licor.

Hacemos una evaluación de varianza constante:

```
plot.ts(diff(ts_data), main = " Serie original diferenciada",
ylab = expression(Z[t]-Z[t-1]), xlab = "time",col='blue',
lwd = 0.1, type = "o")
```



Notamos que no hay una varianza constante, por lo que evaluamos una posible transformación de los datos, así:

```
tBoxCox <- powerTransform(ts_data)
summary(tBoxCox)
```

Obtenemos la salida:

```
bcPower Transformation to Normality
      Est Power Rounded Pwr Wald Lwr Bnd Wald Up Bnd
ts_data  -0.1398           0  -0.6928      0.4132
```

Likelihood ratio test that transformation parameter is equal to 0  
(log transformation)

Likelihood ratio test that no transformation is needed

```
#####
```

```
BoxCox.lambda(ts_data, method=c("loglik"), lower=-2, upper=2)
```

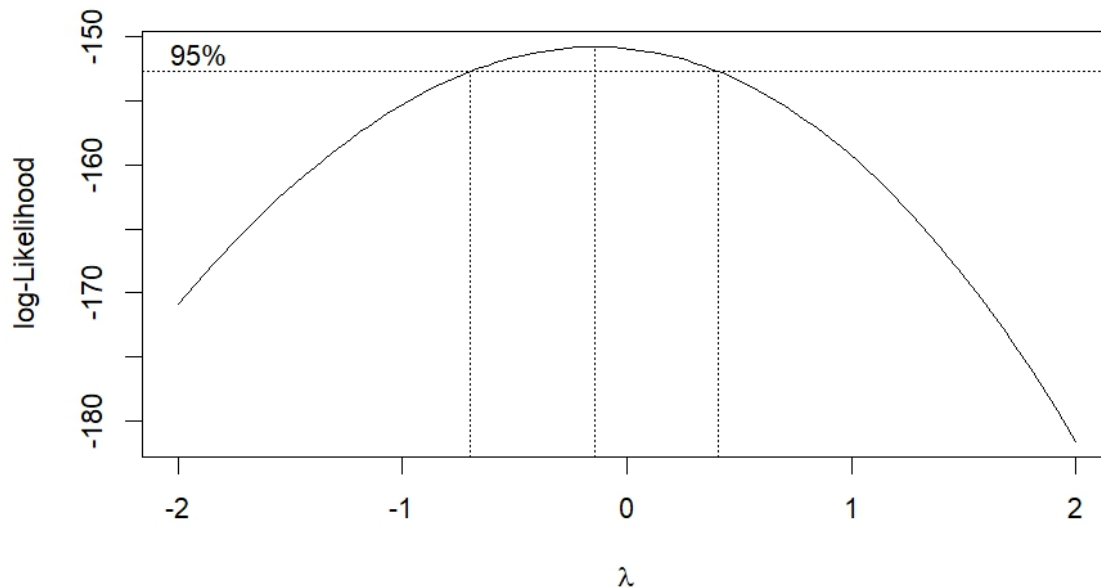
```
[1] 0.15
```

```
BoxCox.lambda(ts_data, method=c("guerrero"))
```

```
[1] -0.1296155
```

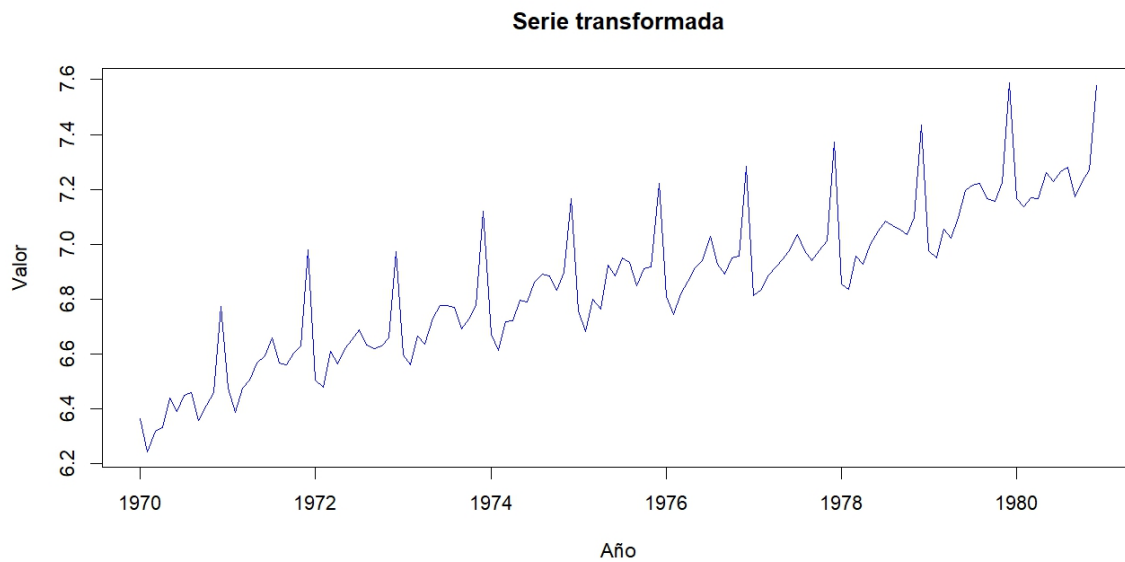
```
# Luego:
```

```
bc <- boxcox(lm(ts_data~1))
```

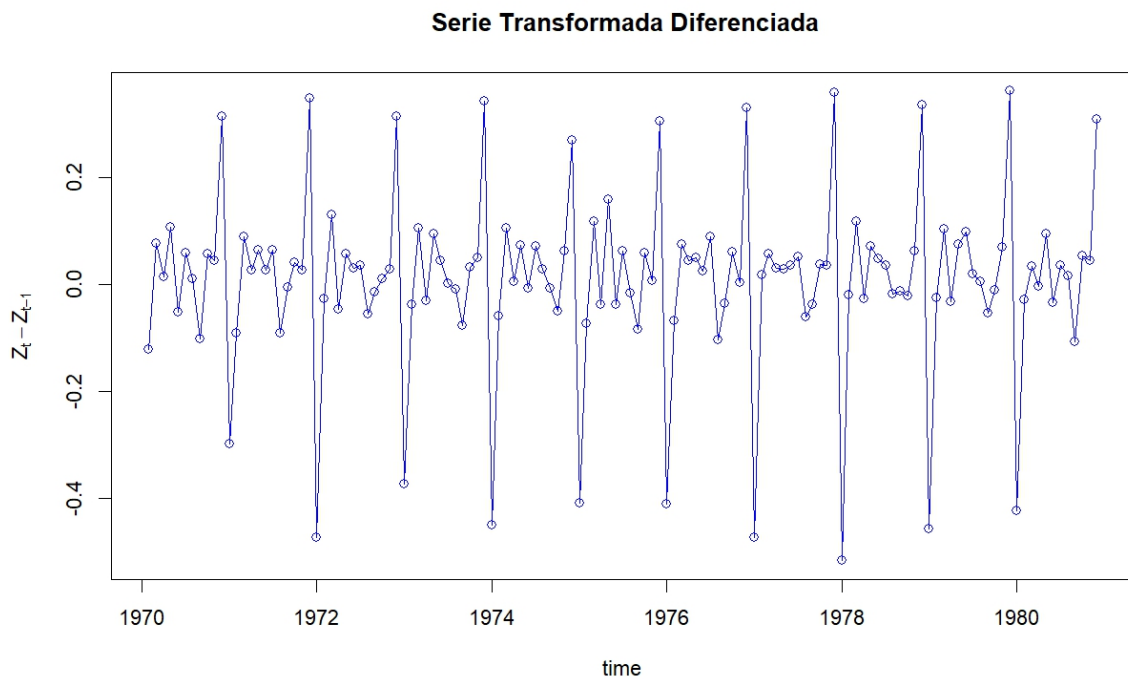


Considerando los métodos aplicados, podemos centrarnos en los candidatos con  $\lambda = 0.15$ ,  $\lambda = -0.12$  y  $\lambda = -0.13$ . Dada la proximidad de estos valores a 0, y al observar la varianza multiplicativa presente en la serie original diferenciada, la mejor opción sería aplicar una transformación logarítmica.

```
plot(log(ts_data), main = "Serie Temporal", xlab = "Año", ylab = "Valor",  
col="blue")
```



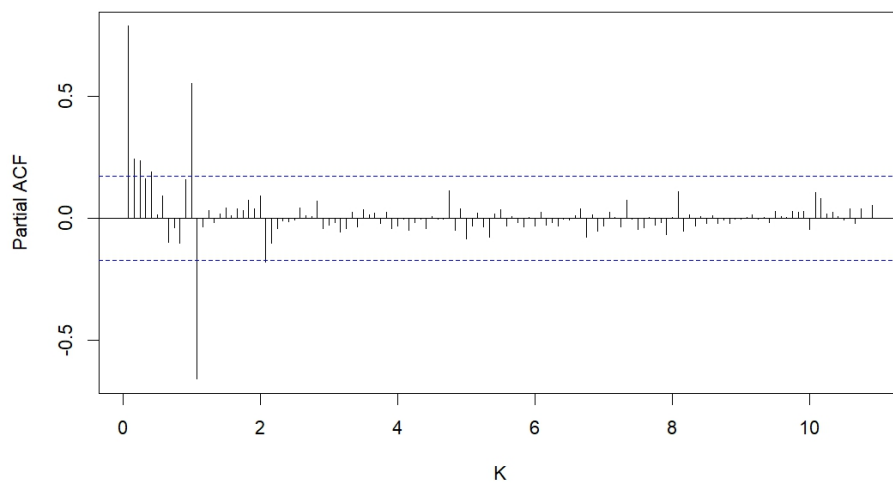
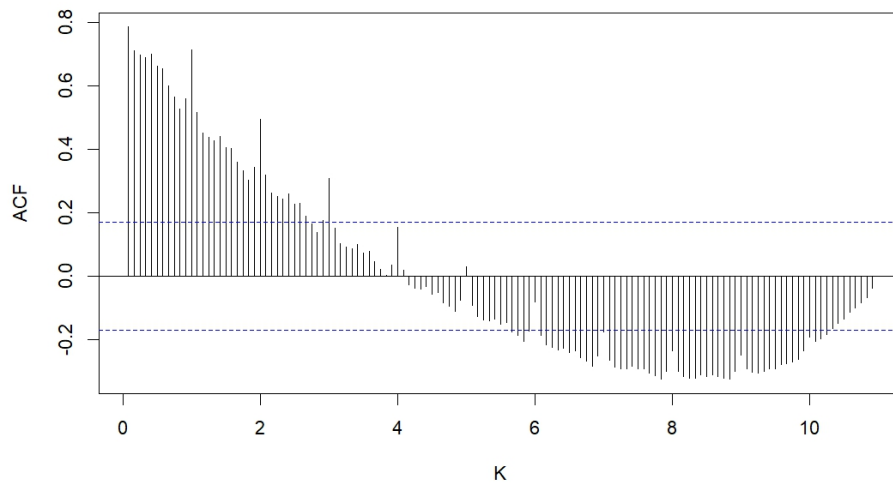
```
plot.ts(diff(log(ts_data)), main = " Serie original diferenciada",
ylab = expression(Z[t]-Z[t-1]), xlab = "time",col='blue',
lwd = 0.1, type = "o")
```



Después de hacer la transformación Box-cot, notamos una mejora en la varianza luego

de hacer la diferenciación a los datos con el logaritmo. Los siguientes análisis se harán bajo la transformación de los datos.

```
Z1 = log(ts_data)
acf(Z1, lag=12, lag.max=132, xlab = "K", main="")
pacf(Z1, lag=12, lag.max=132, xlab = "K", main="")
```



Después de analizar la gráfica de la ACF, se observa un decaimiento geométrico que sugiere la necesidad de revisar la presencia de raíces unitarias en el modelo.

## Prueba de raíces unitarias.

Para realizar la prueba de raíces unitarias se realizara la prueba de Dickey-Fuller, en la que el juego de hipótesis esta dada por:

$$H_0 : \phi = 1$$

$$H_i : \phi < 1$$

Estadístico de prueba:

$$T = \frac{\hat{r}}{\hat{S}_e(\hat{r})}$$

Región de rechazo:

$$T < T_{critico}(\alpha)$$

En esta prueba  $H_0$  establece que la serie tiene una Raíz unitaria, lo que quiere decir que la serie no es estacionaria, por lo tanto rechazar  $H_0$  indicaria que la serie es estacionaria. Para la prueba de Dickey- Fuller vamos a considerar en primera instancia el siguiente modelo:

$$\Delta Z_t = \beta_0 + \beta_1 t + \gamma Z_{t-1} + at$$

Con la ayuda de R tenemos los siguiente coeficientes:

```
# Primera raiz unitaria
```

```
maxlag <- floor(12*(length(Z1)/100)^(1/4)) #Número máximo de rezagos
testDF_1 <- ur.df(Z1, type = c("trend"), lags=maxlag, selectlags = c("BIC"))
summary(testDF_1)
```

Test regression trend

Call:

```
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.116471	-0.030855	-0.004923	0.025680	0.099190

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.318022	0.923310	2.511	0.013594 *
z.lag.1	-0.353975	0.143570	-2.466	0.015318 *
tt	0.002021	0.000873	2.315	0.022551 *
z.diff.lag1	-0.180404	0.149213	-1.209	0.229389
z.diff.lag2	-0.171376	0.139597	-1.228	0.222350
z.diff.lag3	-0.189334	0.129414	-1.463	0.146481

```

z.diff.lag4  -0.209041    0.121192   -1.725  0.087521  .
z.diff.lag5  -0.203580    0.114160   -1.783  0.077457  .
z.diff.lag6  -0.223886    0.107304   -2.086  0.039383  *
z.diff.lag7  -0.261132    0.101114   -2.583  0.011199  *
z.diff.lag8  -0.280731    0.096626   -2.905  0.004483  **
z.diff.lag9  -0.326213    0.091241   -3.575  0.000532  ***
z.diff.lag10 -0.383238    0.087561   -4.377  2.87e-05  ***
z.diff.lag11 -0.408211    0.085400   -4.780  5.79e-06  ***
z.diff.lag12  0.508143    0.083693    6.071  2.09e-08  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.04369 on 104 degrees of freedom
Multiple R-squared:  0.9377,    Adjusted R-squared:  0.9294
F-statistic: 111.9 on 14 and 104 DF,  p-value: < 2.2e-16

```

Value of test-statistic is: -2.4655 8.4871 3.477

Critical values for test statistics:

```

      1pct  5pct 10pct
tau3 -3.99 -3.43 -3.13
phi2  6.22  4.75  4.07
phi3  8.43  6.49  5.47

```

Dado que el intercepto y la tendencia son significativas no hay necesidad de considerar otro modelo, por lo que nos sentramos en los valores critico, en este caso nuestro etadistico de prueba  $T=-2.4655$  asi nuestra region de rechazo en cada nivel de significacia queda de la forma:

	1 %	5 %	10 %
$T < T_{critico}(\alpha)$	-2.46551 <-3.99	-2.4655<-3.43	-2.4655 <-3.13
$T < T_{critico}(\alpha)$	Se acepta	Se acepta	Se acepta

Como se evidencia en la tabla en todos los niveles de significacion se acepta lo hipotesis nula por lo que los datos no son estacionarios y hay la necesidad de diferencia

### Verificacion de segunda raiz unitaria

Para la verificacion de una segunda raiz unitaria consideremos el siguiente modelo

$$\Delta Z_t = \beta_0 + \gamma Z_{t-1} + at$$

Ya que cuando se diferencia se elimita el termino de la tendencia

```

testDF_2 <- ur.df((diff(Z1)), type = c("drift"), lags=maxlag,
selectlags = c("BIC"))
summary(testDF_2)

```

Test regression drift

Call:

```
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.125764	-0.029428	-0.001301	0.031533	0.095820

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.023190	0.008249	2.811	0.005902	**
z.lag.1	-3.539138	1.037151	-3.412	0.000919	***
z.diff.lag1	1.917799	0.998015	1.922	0.057392	.
z.diff.lag2	1.592709	0.913851	1.743	0.084315	.
z.diff.lag3	1.289408	0.829632	1.554	0.123176	
z.diff.lag4	0.985458	0.746888	1.319	0.189927	
z.diff.lag5	0.714744	0.664064	1.076	0.284275	
z.diff.lag6	0.450611	0.581169	0.775	0.439890	
z.diff.lag7	0.178654	0.498102	0.359	0.720570	
z.diff.lag8	-0.086381	0.414431	-0.208	0.835298	
z.diff.lag9	-0.356903	0.329271	-1.084	0.280907	
z.diff.lag10	-0.654592	0.244976	-2.672	0.008753	**
z.diff.lag11	-0.942092	0.163354	-5.767	8.35e-08	***
z.diff.lag12	-0.278811	0.096254	-2.897	0.004599	**

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04334 on 104 degrees of freedom

Multiple R-squared: 0.9771, Adjusted R-squared: 0.9743

F-statistic: 341.9 on 13 and 104 DF, p-value: < 2.2e-16

Value of test-statistic is: -3.4124 5.8873

Critical values for test statistics:

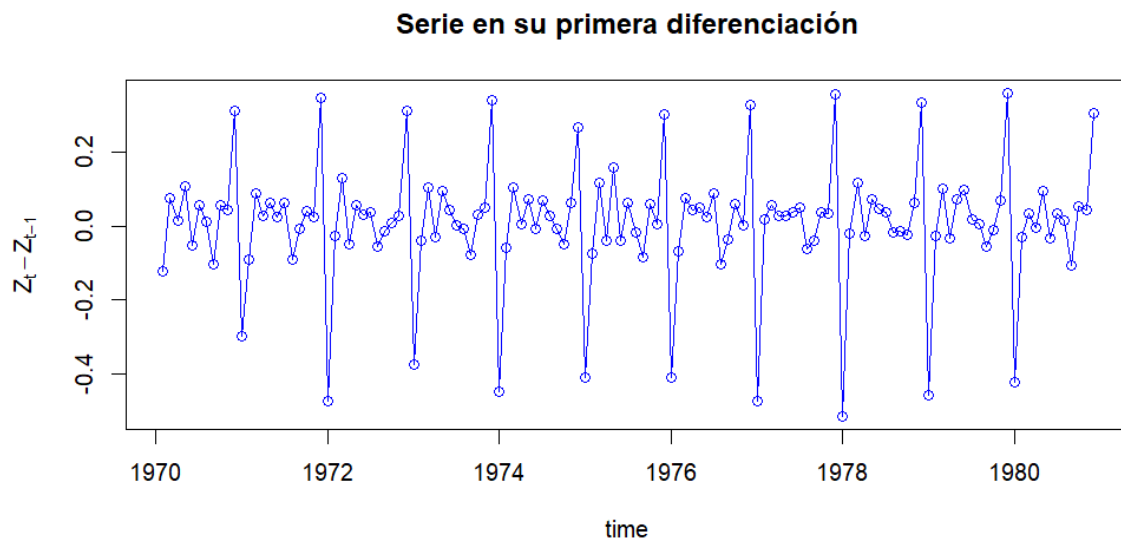
	1pct	5pct	10pct
tau2	-3.46	-2.88	-2.57
phi1	6.52	4.63	3.81

Dado que el intercepto es significativo no hay necesidad de considerar otro modelo, por lo que nos sentramos en los valores critico, en este caso nuestro estadistico de prueba  $T = -3.4124$  asi nuestra region de rechazo en cada nivel de significancia queda de la forma:

	1 %	5 %	10 %
$T < T_{critico}(\alpha)$	-3.4124 < -3.46	-3.4124 < -2.88	-3.4124 < -2.57
$T < T_{critico}(\alpha)$	Se acepta	Se Rechaza	Se rechaza

En este caso, la hipótesis nula se acepta únicamente al nivel de significancia del 1 %, lo que sugiere que la serie no es estacionaria y podría tener una segunda raíz unitaria. Sin embargo, en los niveles de significancia del 5 % y 10 %, se rechaza la hipótesis nula, lo que indica que el proceso es estacionario, es decir, la serie se comporta como estacionaria después de la primera diferenciación. Nos encontramos con dos posibles interpretaciones: seguir el resultado del nivel de significancia del 1 % y buscar una tercera raíz unitaria, o concluir que la serie tiene solo una raíz unitaria. Para tomar la decisión final, examinaremos la gráfica de la serie después de la primera diferenciación.

```
plot.ts(diff(Z1)),lwd = 0.5, type = "o",
ylab = expression(Z[t]-Z[t-1]), xlab = "time",col='blue',
main="Serie en su primera diferenciación")
```



Al observar la serie en su primera diferenciación, representada en la gráfica, se puede notar un comportamiento más estable y estacionario en comparación con la serie original. Este patrón sugiere que la primera diferenciación es efectiva para eliminar la tendencia o la no estacionariedad presente en la serie original.

Dado que los resultados de las pruebas de significancia al 5 % y 10 % rechazan la hipótesis nula, podemos concluir con confianza que la serie es estacionaria después de aplicar la primera diferenciación. Esto se evidencia en la variabilidad relativamente constante alrededor de un valor medio cercano a cero en la gráfica.

Además, se considerará una tercera raíz unitaria al nivel de significancia del 1 %, para evaluar si se requiere una diferenciación adicional o si la serie ya ha alcanzado la estacionariedad deseada en esta etapa.

**Observación:** Sin embargo, también se nota la presencia de picos recurrentes al mismo nivel en cada periodo estacional. Este patrón indica que, aunque la serie parece estacio-



naria en términos regulares, podría ser necesario aplicar una diferenciación estacional adicional para capturar completamente la estacionalidad inherente a los datos.

**Verificación de tercera raíz** Para la verificación de una tercera raíz unitaria consideremos el siguiente modelo

$$\Delta\gamma Z_{t-1} + at$$

Ya que cuando se vuelve a diferencia se elimina el término de el intercepto

```
testDF_3 <- ur.df(diff(diff(Z1)), type = c("none"), lags=maxlag,
selectlags = c("BIC"))
summary(testDF_3)
```

Test regression none

Call:

```
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.160596	-0.028683	0.000974	0.026735	0.112704

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
z.lag.1	-22.63674	2.07512	-10.909	< 2e-16 ***
z.diff.lag1	20.03462	2.00163	10.009	< 2e-16 ***
z.diff.lag2	18.12016	1.85090	9.790	< 2e-16 ***
z.diff.lag3	16.19421	1.69211	9.570	6.14e-16 ***
z.diff.lag4	14.26267	1.52539	9.350	1.90e-15 ***
z.diff.lag5	12.34757	1.34837	9.157	5.12e-15 ***
z.diff.lag6	10.45485	1.16394	8.982	1.26e-14 ***
z.diff.lag7	8.57420	0.97554	8.789	3.37e-14 ***
z.diff.lag8	6.71313	0.78631	8.538	1.21e-13 ***
z.diff.lag9	4.86608	0.60011	8.109	1.07e-12 ***
z.diff.lag10	3.01642	0.42180	7.151	1.24e-10 ***
z.diff.lag11	1.17611	0.25172	4.672	8.95e-06 ***
z.diff.lag12	0.30286	0.09576	3.163	0.00205 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04363 on 104 degrees of freedom

Multiple R-squared: 0.9926, Adjusted R-squared: 0.9917

F-statistic: 1070 on 13 and 104 DF, p-value: < 2.2e-16

Value of test-statistic is: -10.9086

Critical values for test statistics:

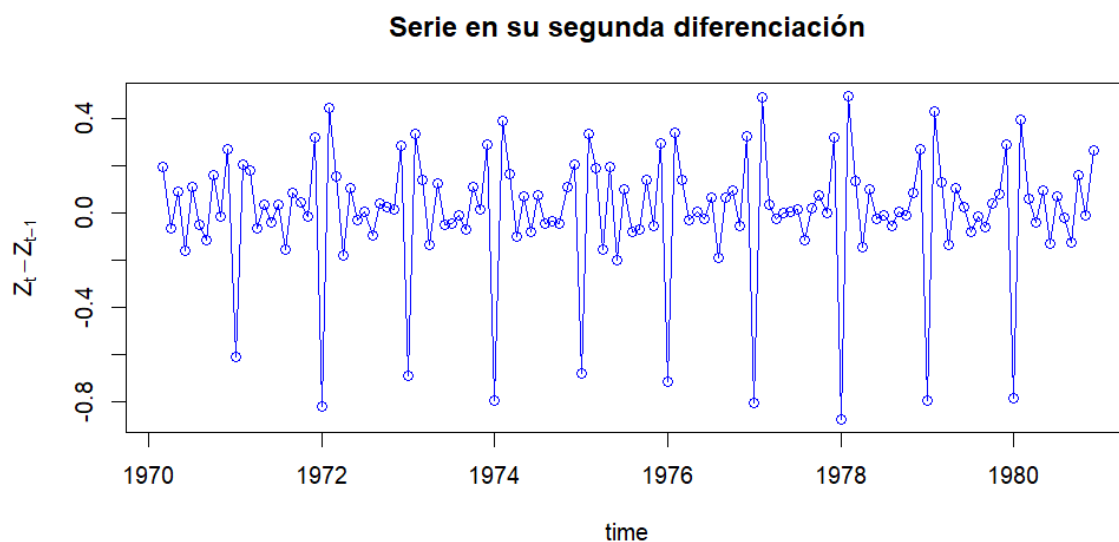
```
1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

Finalmente, luego de la segunda diferenciación obtenemos los siguientes resultados:

	1 %	5 %	10 %
$T < T_{critico}(\alpha)$	-10.9086 < -2.58	-10.9086 < -1.95	-10.9086 < -1.62
$T < T_{critico}(\alpha)$	Se Rechaza	Se Rechaza	Se rechaza

Donde en todos los niveles de significancia, se rechaza la hipótesis nula, por lo que podemos concluir que no hay una tercera raíz unitario lo que significa que la serie en su segunda diferenciación se comporta como un proceso estacionario.

```
plot.ts(diff(diff((Z1))),lwd = 0.5, type = "o",
ylab = expression(Z[t]-Z[t-1]), xlab = "time",col='blue', main="Serie en su segunda di
```



La gráfica correspondiente a la segunda diferenciación muestra un proceso que podría considerarse estacionario. Sin embargo, visualmente, no se observa una mejora significativa en comparación con la serie que resultó de aplicar una sola raíz unitaria. De hecho, la serie después de la primera diferenciación parece representar mejor el comportamiento deseado, lo que sugiere que una segunda diferenciación podría no ser necesaria.

Es importante destacar que la segunda diferenciación se asocia con un nivel de significancia del 1 %. Esto podría llevarnos al riesgo de cometer un error de tipo II, es decir, no rechazar una hipótesis nula cuando en realidad deberíamos hacerlo, al aplicar un criterio demasiado estricto. Además, al observar la estacionalidad, se puede notar

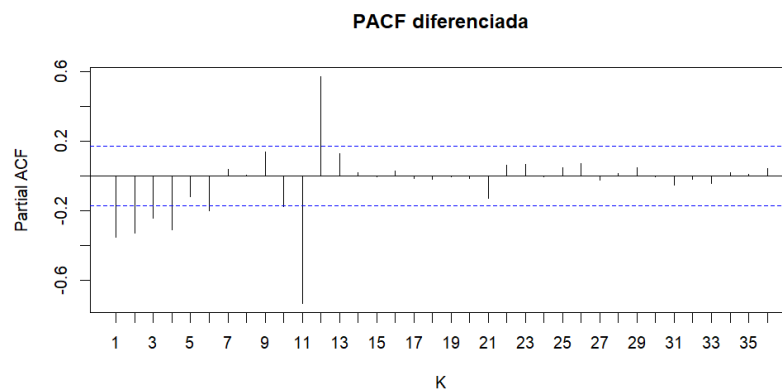
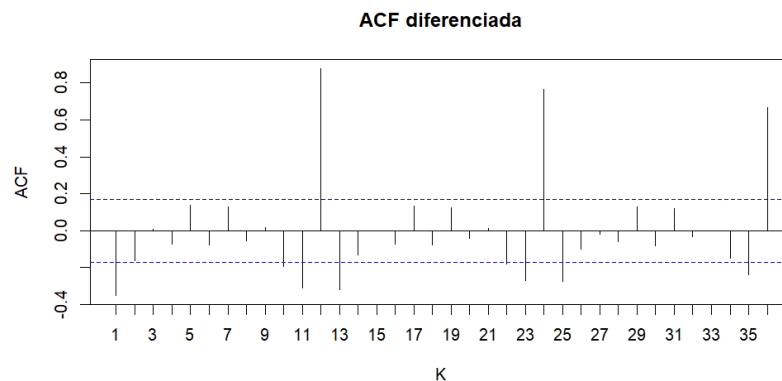
que en cada periodo, la media de la serie tiende a aumentar, lo que podría indicar que la serie se está sobrediferenciando, haciendo que pierda la estacionalidad inherente sin aportar una mejora significativa en la estacionariedad general.

Por lo tanto, considerando tanto el análisis visual como el riesgo asociado con un nivel de significancia demasiado estricto, se recomienda utilizar la serie resultante de la primera diferenciación, ya que parece capturar mejor las características esenciales sin introducir complicaciones innecesarias.

**Conclusión:** La serie presenta una raíz unitaria, lo que indica que es necesario aplicar una diferenciación regular una vez para alcanzar la estacionariedad. Además, se observa que podría ser necesario considerar una diferenciación estacional adicional, debido a los patrones estacionales detectados en la serie.

### ACF y PACF regulares y diferenciadas

```
acf(as.numeric((diff(Z1))), lag.max=36, xlab="K",  
main="ACF diferenciada", xaxp=c(1,36,35))  
pacf(as.numeric((diff(Z1))), lag.max=36, xlab="K",  
main="PACF diferenciada", xaxp=c(1,36,35))
```

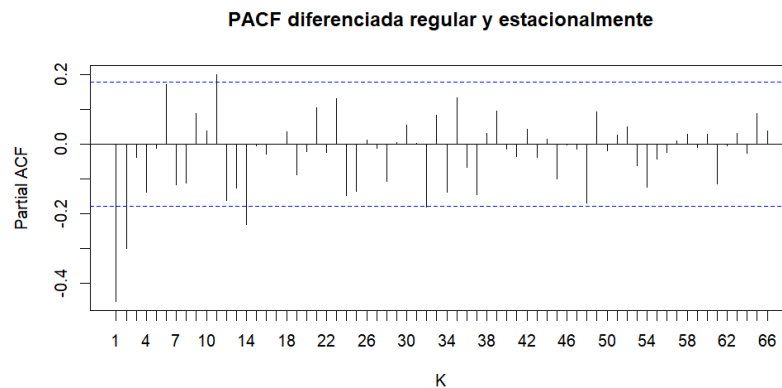
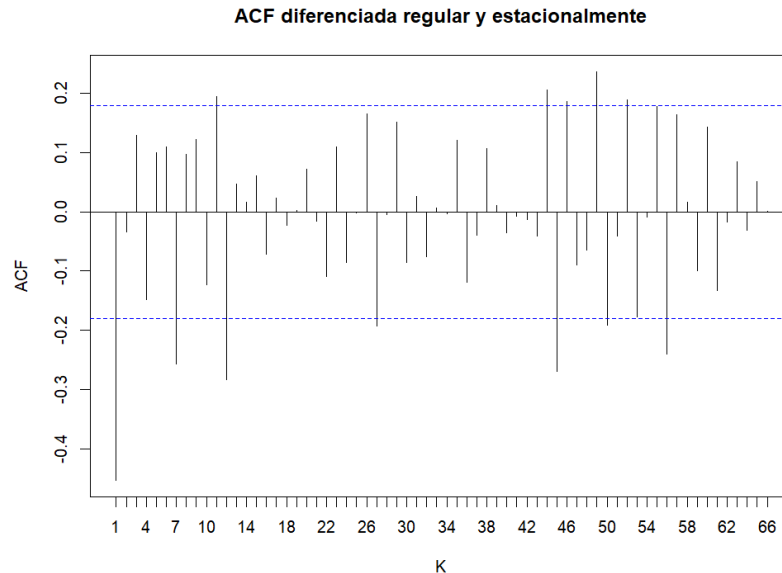


Como se observa en la ACF regular, los periodos estacionales con "S=12" muestran un decaimiento geométrico, lo que indica claramente la necesidad de aplicar una diferenciación estacional para capturar adecuadamente la estructura de la serie. Este hallazgo coincide con los indicios observados previamente en el análisis de raíces unitarias, donde las gráficas sugerían la necesidad de realizar una diferenciación estacional.

### ACF y PACF diferenciadas regulares y estacionales

```
acf(as.numeric(diff(diff(Z1), lag=12)), lag.max=66, xlab="K",
main="ACF diferenciada regular y estacionalmente", xaxp=c(1,66,65))
```

```
pacf(as.numeric(diff(diff(Z1), lag=12)), lag.max=66, xlab="K",
main="PACF diferenciada regular y estacionalmente", xaxp=c(1,66,65))
```



Las conclusiones estarán en el literal B para complementar la elección del modelo.

## 2.2. Literal B)

La idea es identificar un modelo ARIMA estacional. Separamos los datos del último año para la evluación de pronósticos, así:

```
pronosticos <- 12 # Observaciones
pronost_ajutado <- length(Z1)-pronosticos

#Serie omitiendo el último año

Datos_ajuste <- ts(Z1[1:pronost_ajutado], start = c(1970, 1), freq = 12)

# Datos para pronosticar
Datos_pronostico <- ts(Z1[(pronost_ajutado+1):length(Z1)], start=c(1980,1),
                        freq=12)
```

### ARIMA Regular

El ACF muestra un comportamiento de cola, mientras que el PACF presenta cortes en el rezago 2, lo que sugiere la presencia de un componente AR(2). Esto implica que la parte MA es insignificante, es decir, MA(0). Además, dado que se realizaron dos diferenciaciones, tenemos I(1). Por lo tanto, el modelo regular es un ARIMA(2,1,0).

### ARIMA Estacional

La PACF muestra un comportamiento tipo cola en los periodos estacionales S= 12, lo que sugiere que no hay un componente AR estacional significativo. Por otro lado, la ACF presenta un rezago significativo en el periodo estacional 12, lo que indica un componente MA(1). Dado que se aplicó una diferenciación estacional una vez, el modelo estacional queda especificado como ARIMA(0,1,1)

### Modelo propuesto

ARIMA Estacional (2, 1, 0)x(0, 1, 1)[12]

$$(1 - \phi_1\beta - \phi_2\beta^2)(1 - \beta)Z_t = (1 + \Theta_1\beta^{12})\epsilon_t$$

donde:

1.  $\phi_1$  es el coeficiente autorregresivo de orden 1,
2.  $\phi_2$  es el coeficiente autorregresivo de orden 2,
3.  $\Theta_1$  es el coeficiente de la media móvil estacional de orden 1,

4.  $\beta$  es el operador de rezago,
5.  $(1 - \beta)$  representa la diferenciación regular de orden 1,
6.  $\beta^{12}$  indica un rezago estacional de 12 periodos,
7.  $\epsilon_t$  es el término de error en el tiempo t.

### Que modelo propone la funcion auto.arima

```
auto.arima(Datos_ajuste,max.p=5,max.q=5)
Series: Datos_ajuste
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:
          ma1      sma1
      -0.6179  -0.6731
s.e.    0.0753   0.1164

sigma^2 = 0.001224:  log likelihood = 204.27
AIC=-402.55   AICc=-402.32   BIC=-394.53
```

La función auto.arima identificó un modelo ARIMA estacional (0,1,1)x(0,1,1)[12]. En este caso, coincidimos en la parte estacional; sin embargo, en la parte regular, no coincidimos en los órdenes de AR y MA. Según la función, se está interpretando el PACF como un patrón de cola y el ACF como un patrón de rezago, lo cual, en nuestro análisis, es al contrario. A pesar de ello, coincidimos en el orden de diferenciación.

### Ajuste de modelo identificado

```
Series: Datos_ajuste
ARIMA(2,1,0)(0,1,1)[12]

Coefficients:
          ar1      ar2      sma1
      -0.6372  -0.3524  -0.6940
s.e.    0.0933   0.0931   0.1082

sigma^2 = 0.001202:  log likelihood = 205.44
AIC=-402.88   AICc=-402.49   BIC=-392.19
```

#### Training set error measures:

	ME	RMSE	MAE	MPE	MAPE
Training set	-0.0009143109	0.03227785	0.02360001	-0.0164052	0.3449472

	MASE	ACF1
Training set	0.2838998	-0.04807534

Al reemplazar los coeficientes en el modelo, obtenemos la siguiente expresión:

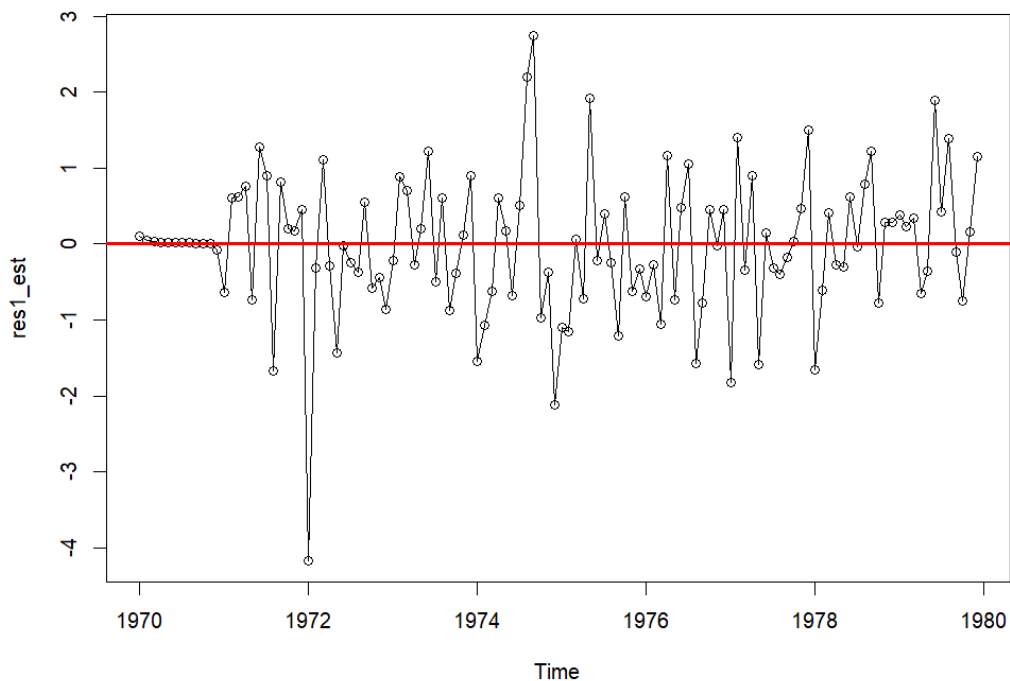
$$(1 + 0.6372\beta + 0.3524\beta^2)(1 - \beta)^2 Z_t = (1 + 0.6940\beta^{12})\epsilon_t$$

### Verificion de los residuos

El objetivo es verificar que los  $\epsilon_t$  se comportan como un ruido blanco gaussiano, donde validamos:

1. Media cero
2. Varianza constante
3. Ningún patrón de dependencia aparente

```
res1_CSS_ML_ML=residuals(modelo)
res1_est=res1_CSS_ML_ML/(modelo$sigma2^.5)
plot.ts(res1_est, type="o")
abline(a=0,b=0,lwd=3, col="red")
```



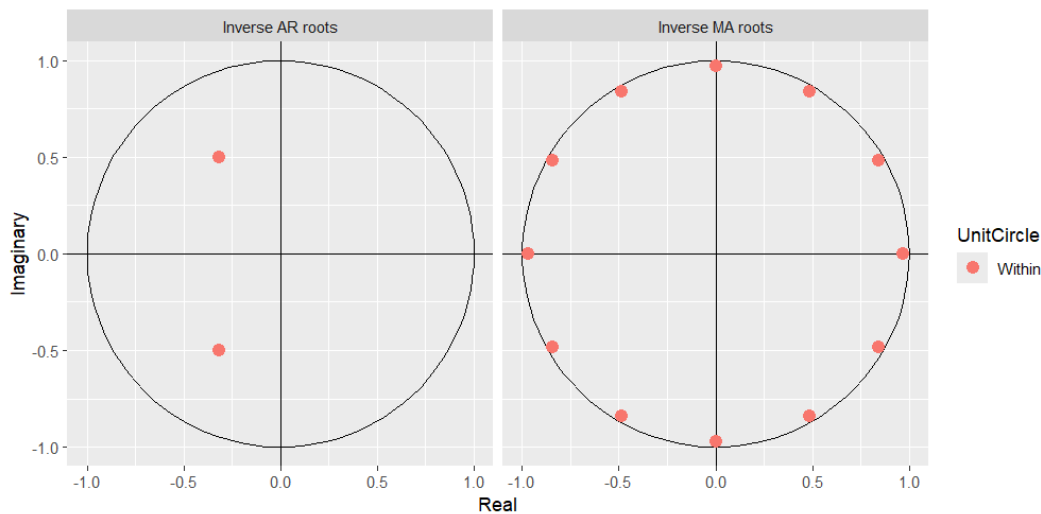
Gráfica de los residuos contra el tiempo

El análisis de los residuos del modelo  $ARIMA(2,1,0) \times (0,1,1)[12]$  muestra que están distribuidos alrededor de cero sin patrones claros de tendencia o estacionalidad, lo que

sugiere que el modelo ha capturado adecuadamente la estructura de los datos. Aunque se observa un pico en los residuos, en general, la varianza parece ser constante a lo largo del tiempo, indicando homocedasticidad. Estos resultados sugieren que los residuos son ruido blanco, lo que refuerza la validez del modelo para predecir los valores de la serie temporal.

También podemos ver que se comporta de forma estacional, para validar esta afirmación, evaluamos las raíces del polinomio utilizando la función `autoplot` en R.

```
autoplot(modelo)
```

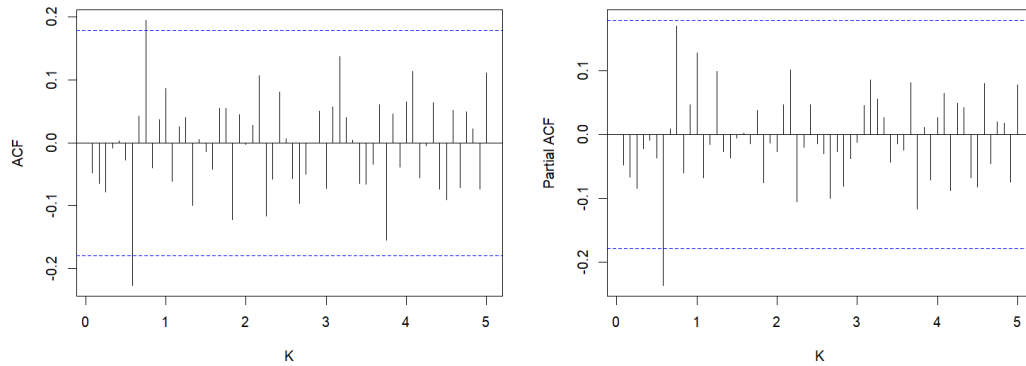


**Gráfica de las raíces inversas**

El análisis de las raíces inversas del modelo  $\text{ARIMA}(2,1,0) \times (0,1,1)[12]$  revela que las 12 raíces correspondientes al componente estacional de media móvil (MA) están dentro del círculo unitario, lo que confirma la invertibilidad del componente estacional. La presencia de estas 12 raíces se debe a la periodicidad estacional del modelo, que considera un ciclo de 12 periodos (por ejemplo, meses en un año). Además, las raíces del componente autoregresivo (AR) también se encuentran dentro del círculo unitario, indicando que el modelo es estable. En conjunto, estos resultados aseguran que el modelo es adecuado para capturar tanto la estructura regular como la estacional de los datos, proporcionando predicciones fiables y robustas.

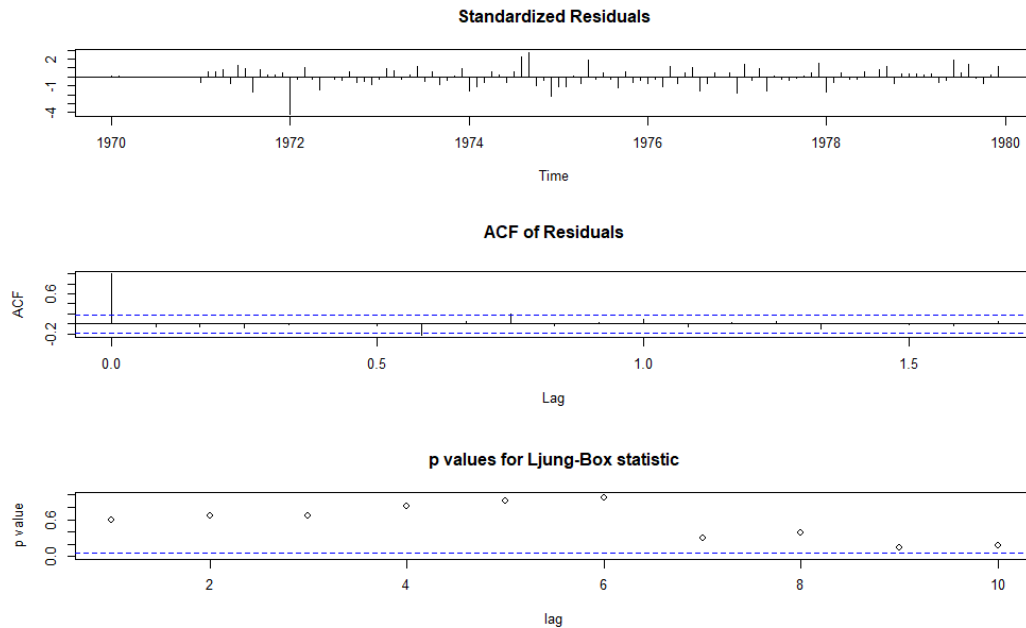
```
par(mfrow = c(1,2))
acf(residuals(modelo), lag.max = 60, xlab = "K", main = "")
pacf(residuals(modelo), lag.max = 60, xlab = "K", main = "")
```





Los gráficos de ACF y PACF de los residuos del modelo  $ARIMA(2,1,0) \times (0,1,1)$  [12] muestran que no hay autocorrelación significativa en la mayoría de los retardos. Esto sugiere que los residuos pueden ser considerados ruido blanco, lo que indica que el modelo ha capturado adecuadamente la estructura de los datos, incluyendo la estacionalidad mensual. En resumen, estos resultados respaldan la validez del modelo ajustado, confirmando que no queda información significativa sin modelar en los residuos.

`tsdiag(modelo)`



Observamos que los residuos estandarizados muestran un comportamiento que se asemeja a una distribución normal, con algún valor atípico. Para identificar un dato como

atípico, consideramos aquellos que superan el umbral de  $(-3, 3)$ . El análisis de los residuos del modelo  $ARIMA(2,1,0) \times (0,1,1)$  [12] sugiere que, en general, los residuos pueden considerarse ruido blanco, como lo confirman los p-valores de la prueba de Ljung-Box y la ACF de los residuos. Se detecta un residuo atípico alrededor de 1972 que sobresale en los residuos estandarizados. Aunque parecer solo un dato atípico, el modelo en su conjunto parece capturar adecuadamente la estructura de la serie temporal analizada.

### Datos atípicos

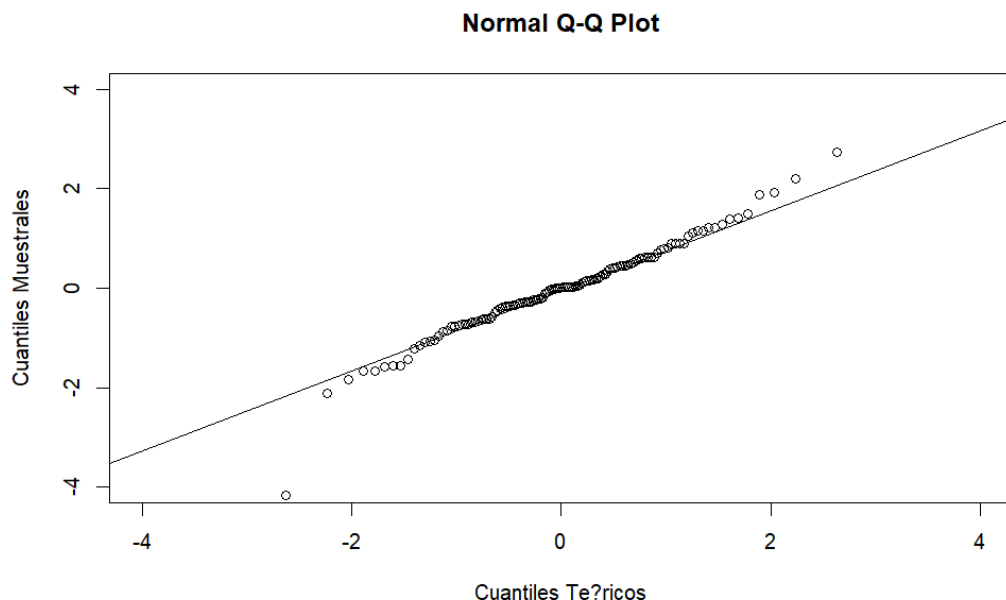
```
(Nobs_Esp=round(length(Datos_ajuste)*2*pnorm(-3, mean = 0, sd = 1,
lower.tail = TRUE)))
[1] 0
```

```
ind=(abs(res1_est)>3.0)
sum(ind)
[1] 1
```

Aunque se detectó una observación atípica que supera las 3 desviaciones estándar, lo que podría afectar la normalidad de los residuos, el número de datos atípicos es muy reducido. Por lo tanto, es poco probable que estos pocos valores tengan un impacto significativo en el análisis global o en las suposiciones del modelo.

### Verificacion de normalidad

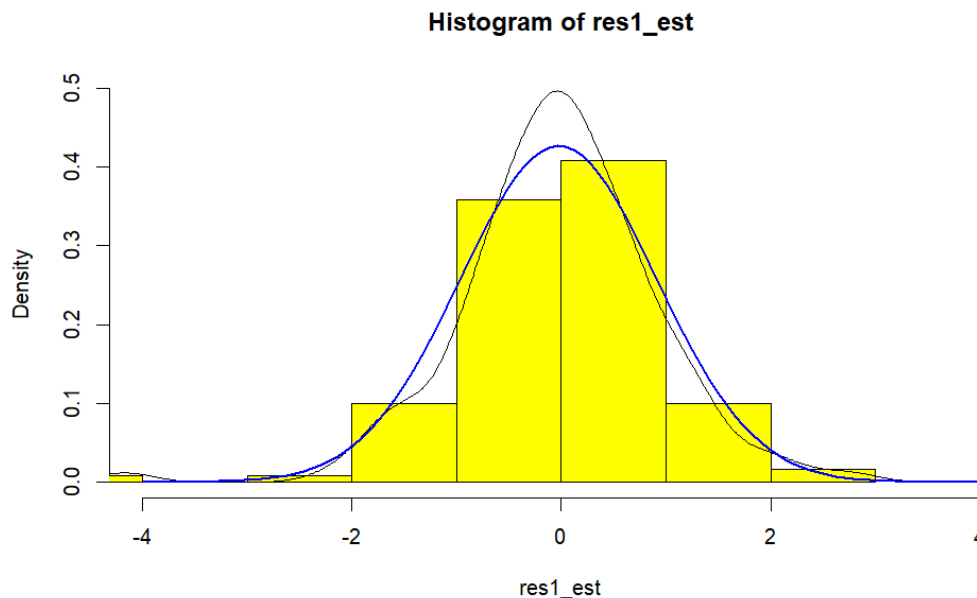
```
qqnorm(res1_est, xlab = "Cuantiles Teóricos", ylab = "Cuantiles Muestrales",
xlim=c(-4,4), ylim=c(-4,4))
qqline(res1_est)
```



```

mu<-mean(res1_est)
sigm<-sd(res1_est)
x<-seq(-4,4,length=100)
y<-dnorm(x,mu,sigm)
hist(res1_est,prob=T,ylim=c(0,0.5),xlim=c(-4,4),col="yellow")
lines(density(res1_est))
lines(x,y,lwd=2,col="blue")

```



```
shapiro.test(res1_est)
```

Shapiro-Wilk normality test

```

data:  res1_est
W = 0.96531, p-value = 0.003459

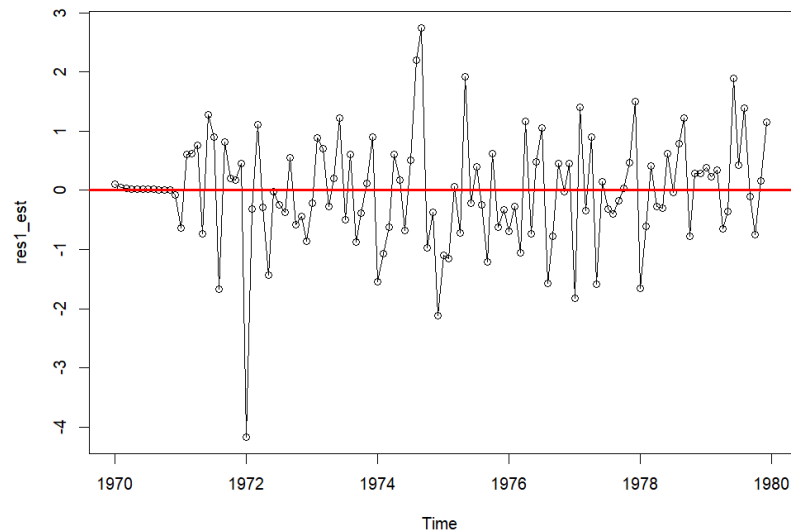
```

Dado que el p-valor es menor que el nivel de significancia del 5 %, rechazamos la hipótesis nula de que los residuos siguen una distribución normal. Esto indica que los residuos no son perfectamente normales, lo que podría ser una señal de que hay algún grado de asimetría o curtosis en los residuos.

### valores ajustados vs los valores originales

Aunque el modelo  $ARIMA(2,1,0) \times (0,1,1)_{12}$  proporciona un buen ajuste visual a la serie temporal, capturando adecuadamente tanto la tendencia como la estacionalidad, el test de Shapiro-Wilk revela que los residuos no siguen una distribución normal. Esto sugiere que, si bien el modelo es efectivo para predecir la serie y refleja bien la estructura de los datos, no se cumplen completamente las suposiciones de normalidad en los

H



residuos. Aunque esta falta de normalidad no invalida el modelo para fines predictivos, podría ser relevante si se requieren supuestos estrictos de normalidad para análisis o inferencias adicionales. Es importante considerar esta limitación al interpretar los resultados y, si es necesario, explorar ajustes adicionales o transformaciones para mejorar el cumplimiento de la normalidad.

### Referencias:

Para el primer punto, en el que utilizamos un modelo de regresión lineal con errores estructurales ARMA para implementarlo en R, nos basamos en el siguiente enlace: <https://otexts.com/fpp2/regarima.html>