ANALISIS SERIES DE TIEMPO - RETAIL

José Nicolás Plaza Bastidas

22-06-2022

La serie de tiempo corresponde a las ventas en unidades de una empresa de retail chilena, la cuál contiene información mensual desde 2017 hasta febrero del año 2020. El objetivo es predecir la venta en unidades en un horizonte de 2 meses. Es por ello, que separaremos la data de entrenamiento usando hasta fines del año 2019 y los dos meses del año 2020 como data de prueba.

Las bases de datos de retail son famosas por presentar una estacionalidad bien marcada. La idea es aprovechar este comportamiento para estimar las unidades que podrían venderse y así tomar decisiones logísticas para evitar que exista agotamiento de stock para las fechas de mayor demanda.

1.- Análisis descriptivo:

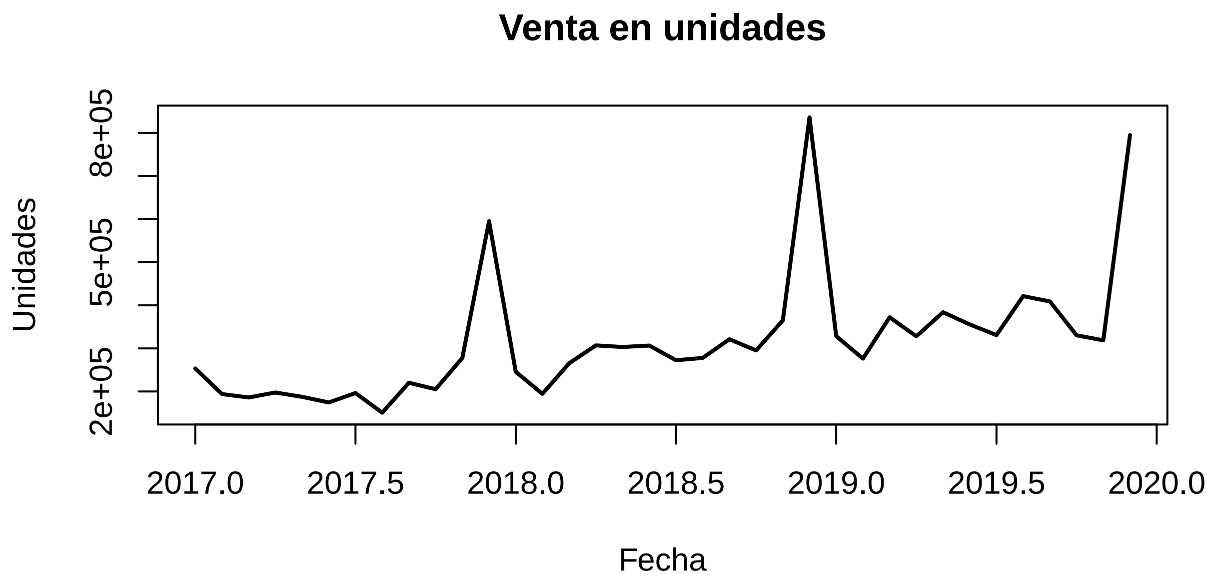
Un análisis descriptivo de los datos se muestra a continuación: **Visualización de los datos**

```
##         Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
## 2017 253444 193992 186097 197583 187599 174542 196361 150744 220241 205199
## 2018 245732 194537 265426 307054 303288 306632 272489 277905 321273 295376
## 2019 328378 276340 372333 328097 384015 355664 330853 421421 409064 330470
##         Nov    Dec
## 2017 277863 595356
## 2018 364986 836414
## 2019 318822 795181
```

Summary

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  150744  216481  299332  318910  337056  836414
```

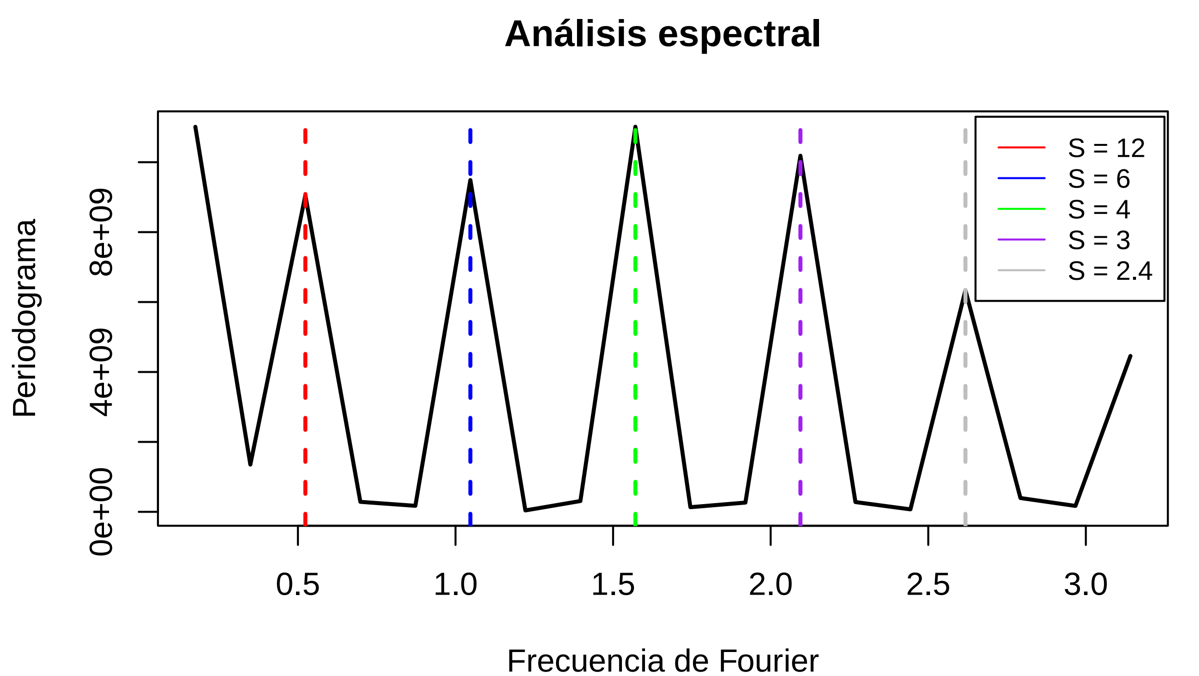
Podemos observar que el valor máximo se encuentra muy lejano a los demás cuartiles. Esto se debe al efecto producido en el mes de diciembre por la navidad, donde hay un incremento sustancial en las ventas. Esto se puede ver de mejor forma en el gráfico.



Se puede ver que la serie tiene una leve tendencia y períodos de gran venta en las navidades.

Análisis espectral

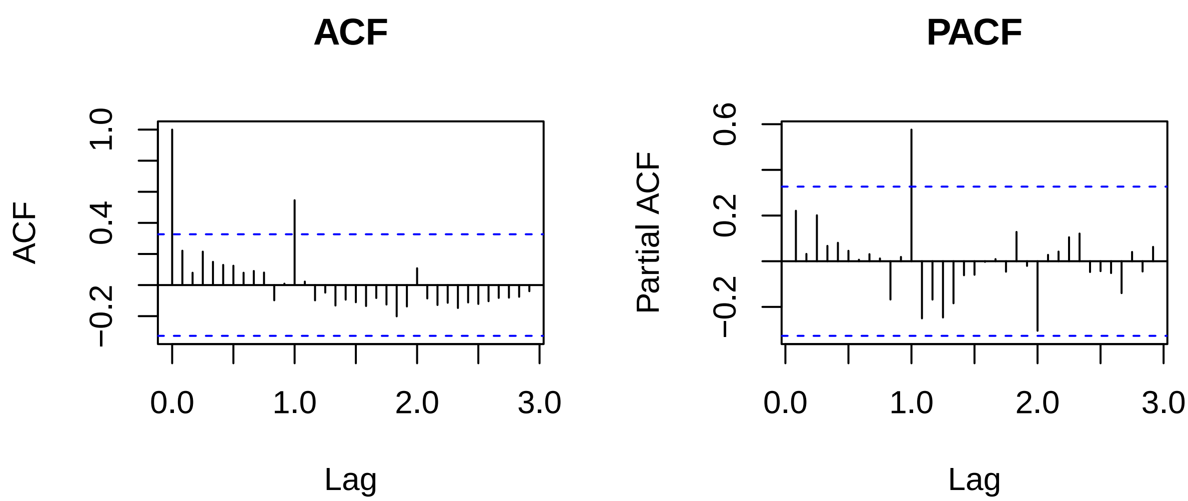
Se realizó un análisis espectral mediante un periodograma para poder detectar estacionalidad en los datos, para así saber cuales son los períodos estacionales que predominan.



Este gráfico también nos otorga información para saber si la serie debe ser diferenciada o no para quitar el efecto de la tendencia. Si la tendencia fuera alta, en el periodograma no podríamos distinguir los períodos estacionales y solo se vería una linea curva apegada a los ejes X e Y.

En este caso, el periodograma si nos muestra de forma clara los períodos estacionales más fuertes. El problema es que tienen todos una altura semejante en el gráfico por lo cual es dificil tomar una decisión para saber con qué período quedarnos. Para complementar este análisis, utilizaremos la información del gráfico ACF que veremos a continuación para ver si logramos encontrar patrones estacionales.

2.- Gráficos ACF y PACF



ACF:

- 1 rezago se sale de la banda.
- Se detecta un patrón estacional que ocurre cada 12 meses.
- No se observa decaimiento exponencial.
- Dada la estacionalidad, podemos proponer un MA(1) combinado con SMA(1) a simple vista.

PACF

- 1 rezago se sale de la banda.
- No se logra ver un patrón claro estacional.
- Podemos proponer un AR(1) combinado con SAR(1) a simple vista dado que sabemos que la serie tiene estacionalidad marcada.

Viendo los gráficos ACF, PACF y Periodograma, podemos ver que existe un patrón estacional marcado que se repite cada 12 meses. Lo cual es una información importante al momento de proponer el modelo.

3.- Proposición de modelo de la familia SARIMA(p,d,q)X(P,D,Q)[S]

Con las observaciones de los gráficos vistos, definiremos una grilla para un modelo SARIMA para distintas combinaciones de p,q,P,Q. los cuales tomarán valores 0 y 1. También veremos como se comporta si agregamos una diferenciación estacional (D=1).

Una vez definida la grilla, compararemos los modelos según los criterios AIC y BIC de todos para seleccionar los mejores.

```
t = 12
fit1  <- Arima(y = dt, order = c(1,0,1), seasonal= list(order=c(1,1,1),period=t))
fit2  <- Arima(y = dt, order = c(1,0,1), seasonal= list(order=c(0,1,1),period=t))
fit3  <- Arima(y = dt, order = c(1,0,1), seasonal= list(order=c(1,1,0),period=t))
fit4  <- Arima(y = dt, order = c(1,0,1), seasonal= list(order=c(0,1,0),period=t))
fit5  <- Arima(y = dt, order = c(1,0,0), seasonal= list(order=c(1,1,1),period=t))
fit6  <- Arima(y = dt, order = c(1,0,0), seasonal= list(order=c(0,1,1),period=t))
fit7  <- Arima(y = dt, order = c(1,0,0), seasonal= list(order=c(1,1,0),period=t))
fit8  <- Arima(y = dt, order = c(1,0,0), seasonal= list(order=c(0,1,0),period=t))
fit9  <- Arima(y = dt, order = c(0,0,1), seasonal= list(order=c(1,1,1),period=t))
fit10 <- Arima(y = dt, order = c(0,0,1), seasonal= list(order=c(0,1,1),period=t))
fit11 <- Arima(y = dt, order = c(0,0,1), seasonal= list(order=c(1,1,0),period=t))
fit12 <- Arima(y = dt, order = c(0,0,1), seasonal= list(order=c(0,1,0),period=t))
fit13 <- Arima(y = dt, order = c(0,0,0), seasonal= list(order=c(1,1,1),period=t))
fit14 <- Arima(y = dt, order = c(0,0,0), seasonal= list(order=c(0,1,1),period=t))
fit15 <- Arima(y = dt, order = c(0,0,0), seasonal= list(order=c(1,1,0),period=t))
```

Selección AIC

```
##       df      AIC
## fit8   2 600.8871
## fit4   3 601.0260
## fit7   3 602.6677
## fit6   3 602.6737
## fit2   4 602.9609
## fit3   4 602.9632
## fit5   4 604.6324
## fit1   5 604.9108
## fit12  2 610.9477
## fit11  3 611.4991
## fit10  3 611.5477
## fit9   4 613.3932
## fit14  2 621.0651
## fit15  2 621.0651
## fit13  3 623.0651
```

Según el criterio AIC se seleccionan los modelos

- 1.- SARIMA(1,0,0)(0,1,0)[12] (fit8)
- 2.- SARIMA(1,0,1)(0,1,0)[12] (fit4)
- 3.- SARIMA(1,0,0)(1,1,0)[12] (fit7)

Selección BIC

```
##        df      BIC
## fit8   2 603.2432
## fit4   3 604.5602
## fit7   3 606.2019
## fit6   3 606.2078
## fit2   4 607.6731
## fit3   4 607.6754
## fit5   4 609.3446
## fit1   5 610.8010
## fit12  2 613.3038
## fit11  3 615.0333
## fit10  3 615.0819
## fit9   4 618.1054
## fit14  2 623.4212
## fit15  2 623.4212
## fit13  3 626.5993
```

Según el criterio BIC se seleccionan los modelos

- 1.- SARIMA(1,0,0)(0,1,0)[12] (fit8)
- 2.- SARIMA(1,0,1)(0,1,0)[12] (fit4)
- 3.- SARIMA(1,0,0)(1,1,0)[12] (fit7)

4.- Elección del modelo

En base a lo anterior, se analizarán los mejores tres modelos puntuado según ambos criterios, es decir

- 1.- SARIMA(1,0,0)(0,1,0)[12]

$$(1 - \phi_1 \cdot B)(1 - B^{12})X_t = \epsilon_t$$

donde D = 1, S = 12 y ϕ_1 = 0.7629

- 2.- SARIMA(1,0,1)(0,1,0)[12]

$$(1 - \phi_1 \cdot B)(1 - B^{12})X_t = (1 + \theta_1)\epsilon_t$$

donde D = 1, S = 12, ϕ_1 = 0.8908 y θ_1 = -0.3621

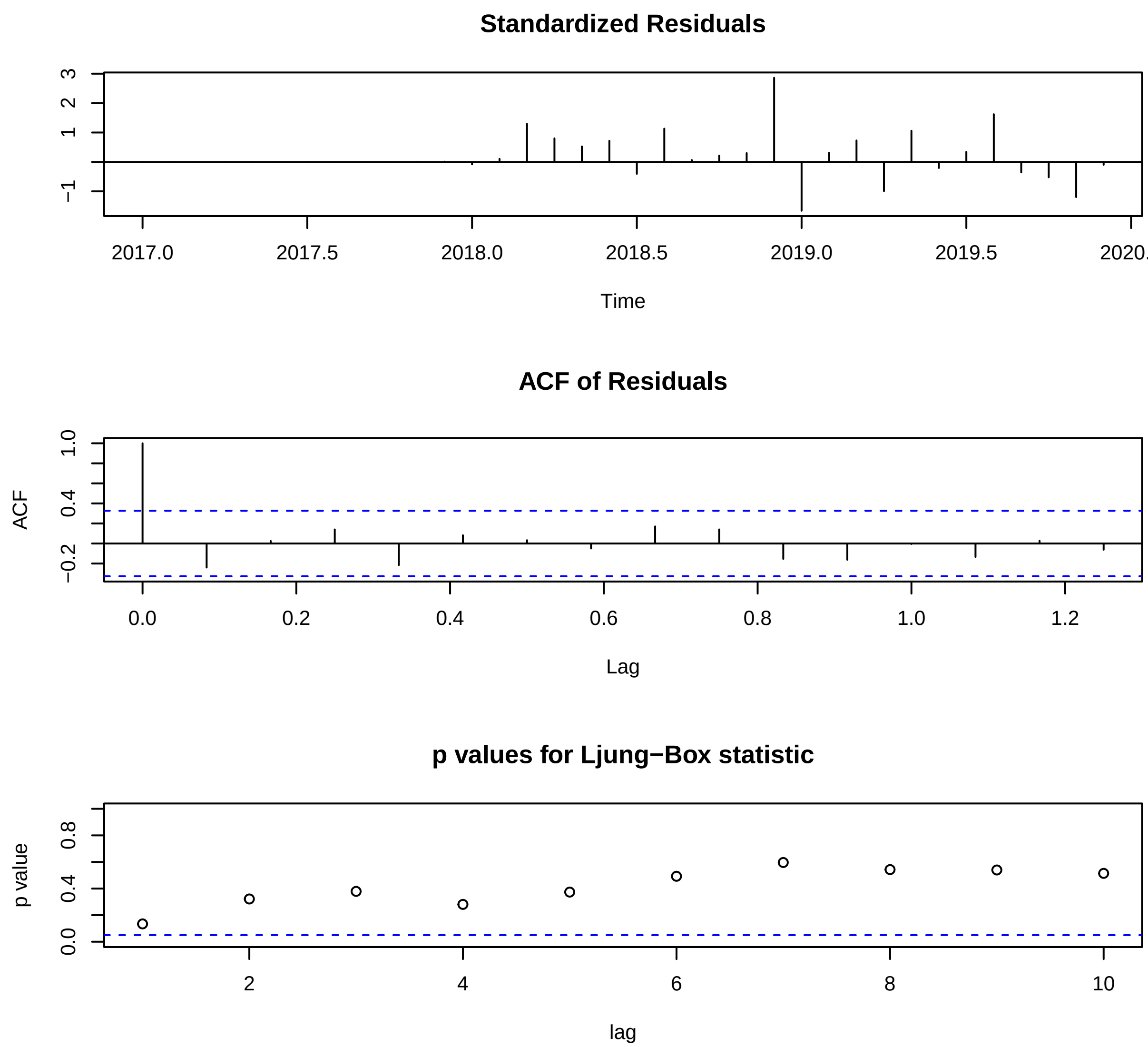
- 3.- SARIMA(1,0,0)(1,1,0)[12]

$$(1 - \phi_1 \cdot B)(1 - \phi_1^{'} \cdot B^{12})(1 - B^{12})X_t = \epsilon_t$$

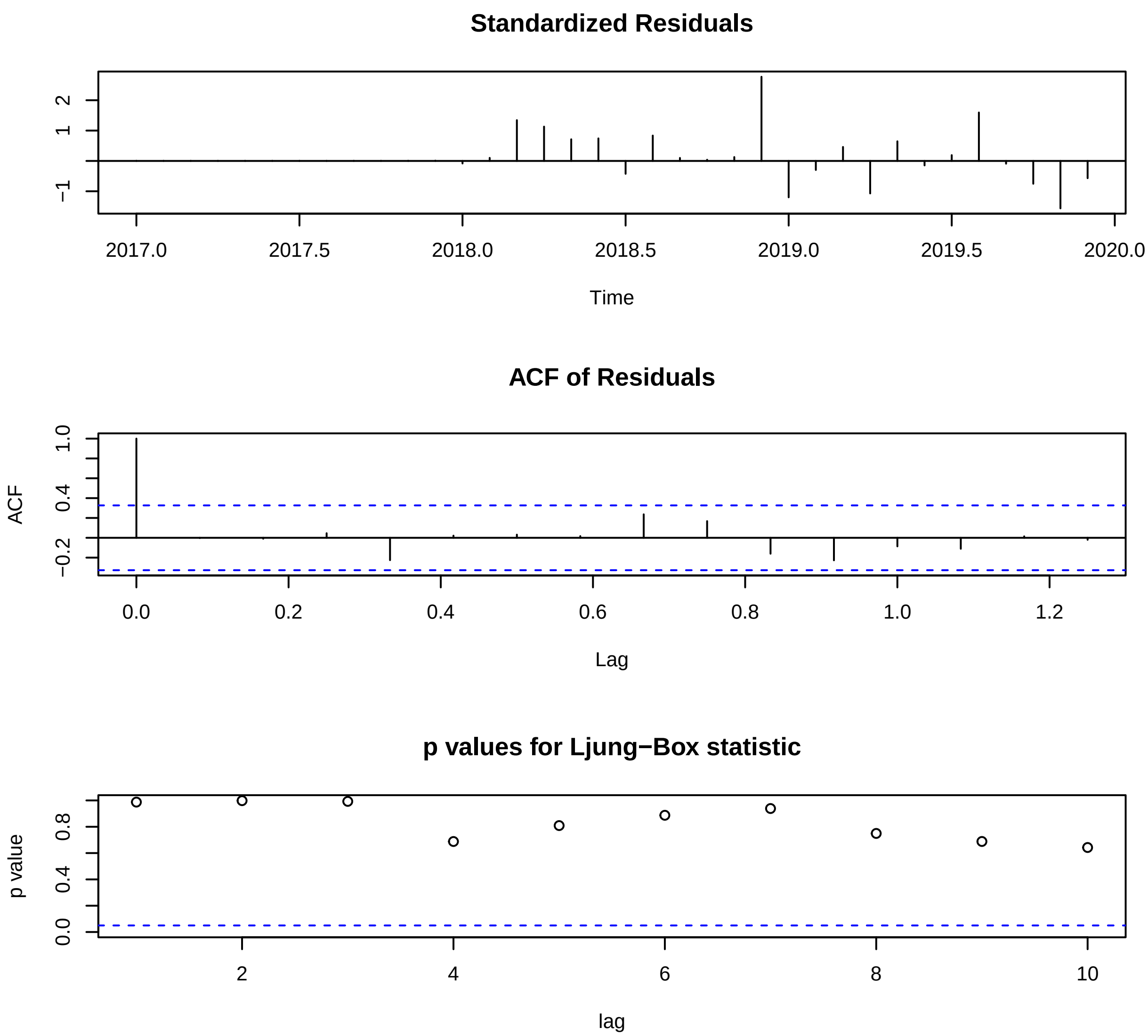
donde D = 1, S = 12, ϕ_1 = 0.7583 y $\phi_1^{'}$ = 0.1474

5.- Evaluación de supuestos

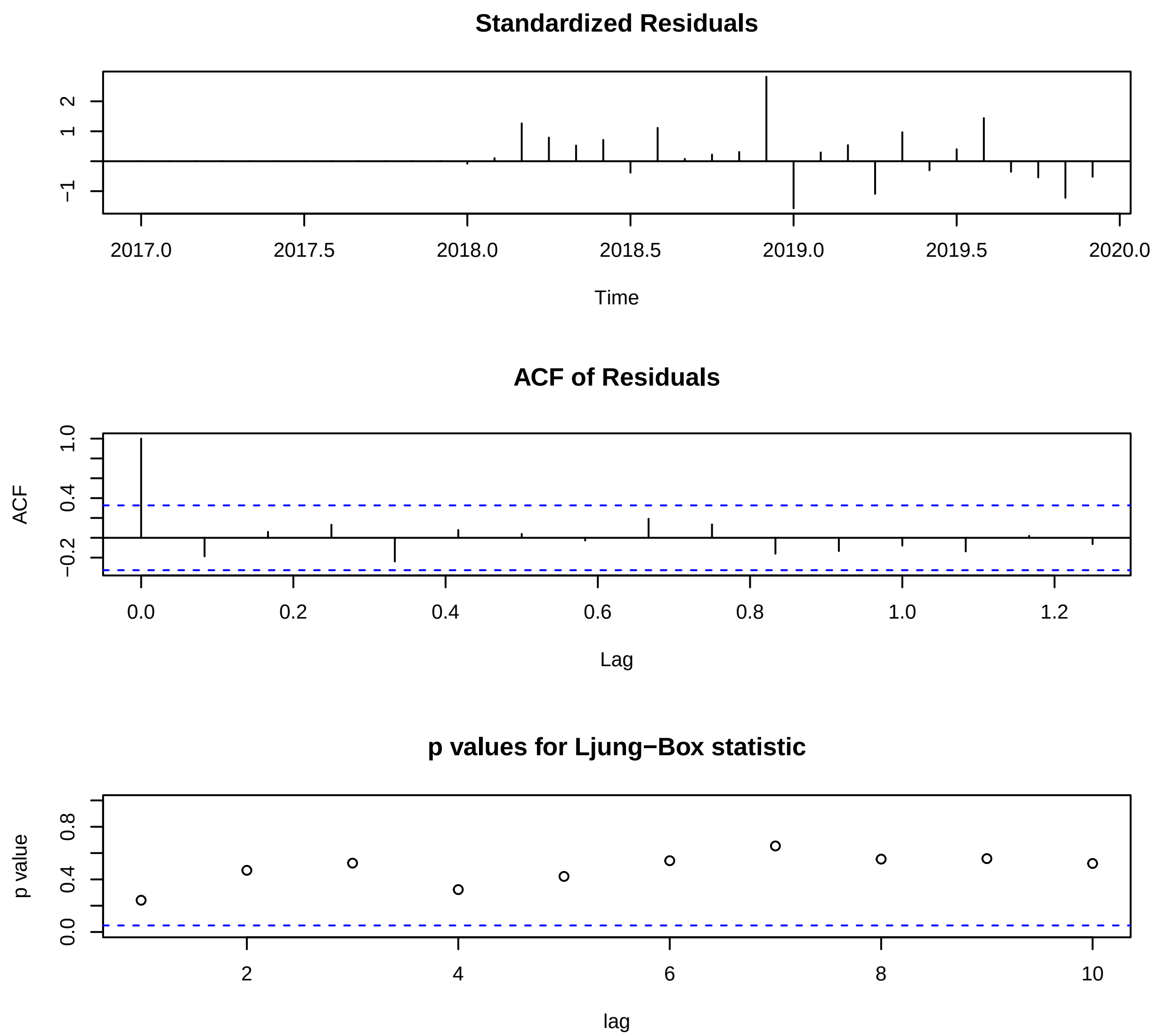
Con la función `tsdiag` podemos observar los p-value y ver si los residuos del modelo son un ruido blanco o no:



Para el modelo SARIMA(1,0,0)(0,1,0)[12], el gráfico Box-Ljung muestra que los p-value están por sobre la banda de confianza, pero en el primer lag se ve un p-value muy abajo, por lo que nos puede dar indicios de que no se cumple muy bien el supuesto de indepencia y podría no ser un ruido blanco.

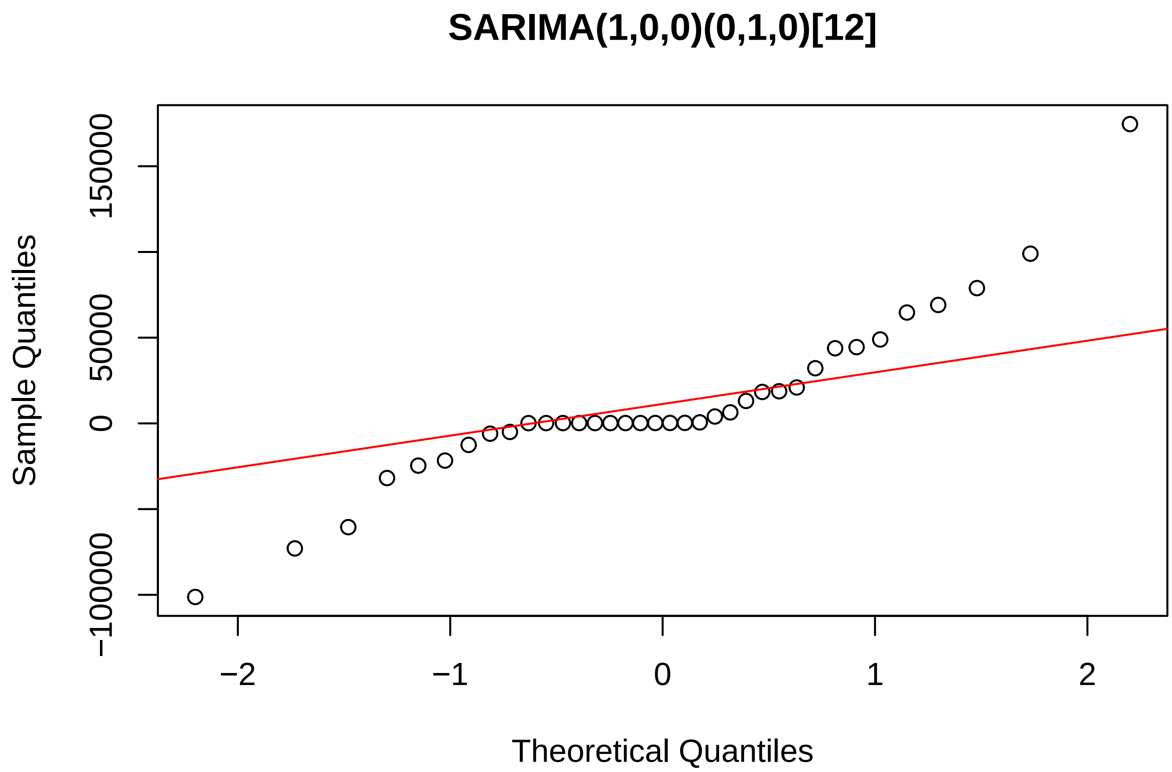


Para el modelo SARIMA(1,0,1)(0,1,0)[12], el gráfico Box-Ljung muestra que los p-value están muy arriba por sobre la banda de confianza, por lo que se cumple que los residuos corresponden a un ruido blanco.



Para el modelo SARIMA(1,0,0)(1,1,0)[12], el gráfico Box-Ljung muestra que los p-value están por sobre la banda de confianza, por lo que se cumple que los residuos corresponden a un ruido blanco.

Normalidad



Graficamente podemos ver que los residuos no siguen una distribución normal. Para complementar este análisis realizaremos un test de Shapiro-Wilks.

Definimos el test de la siguiente forma:

- H_0 : Los residuos se distribuyen normal
- H_1 : Los residuos no se distribuyen normal

Nivel de significancia:

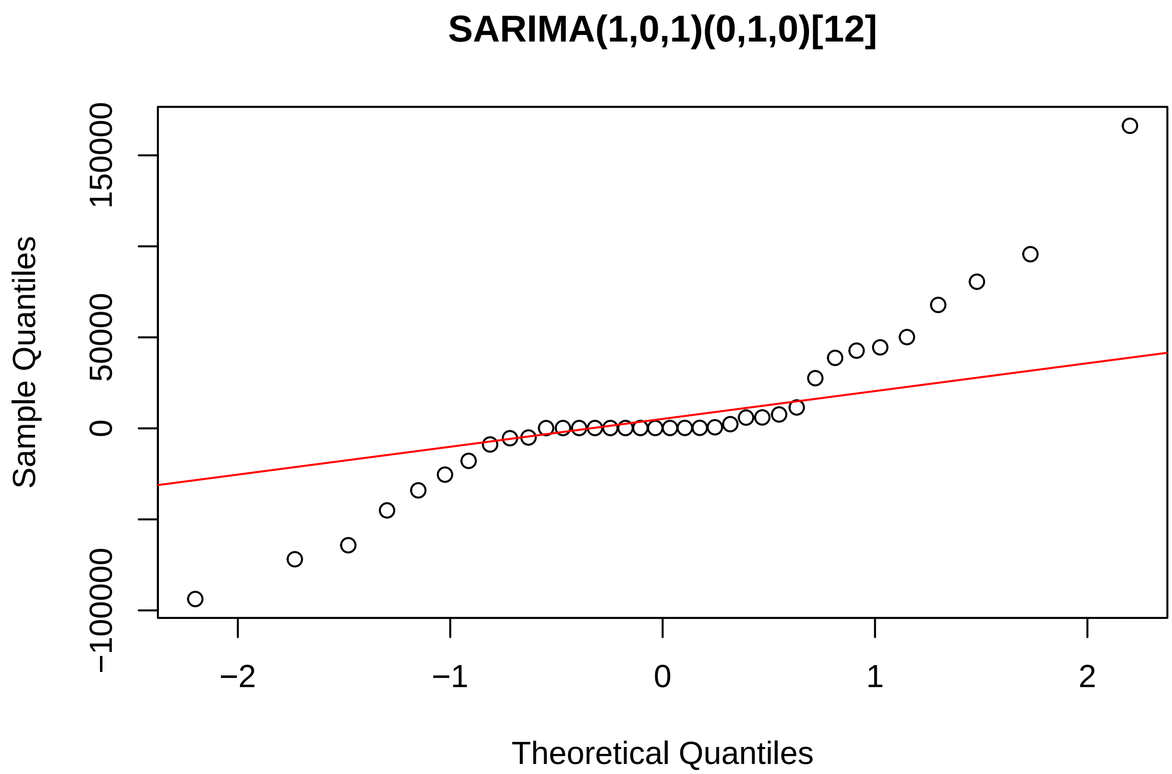
- $\alpha = 5\%$

Rechazamos H_0 si el p-value es menor al nivel de significancia definido.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos1
## W = 0.89323, p-value = 0.00224
```

Vemos que el `p-value` es menor que $\alpha = 5\%$, por lo que rechazamos H_0, es decir, los residuos no siguen una distribución normal para el modelo SARIMA(1,0,0)(0,1,0)[12].

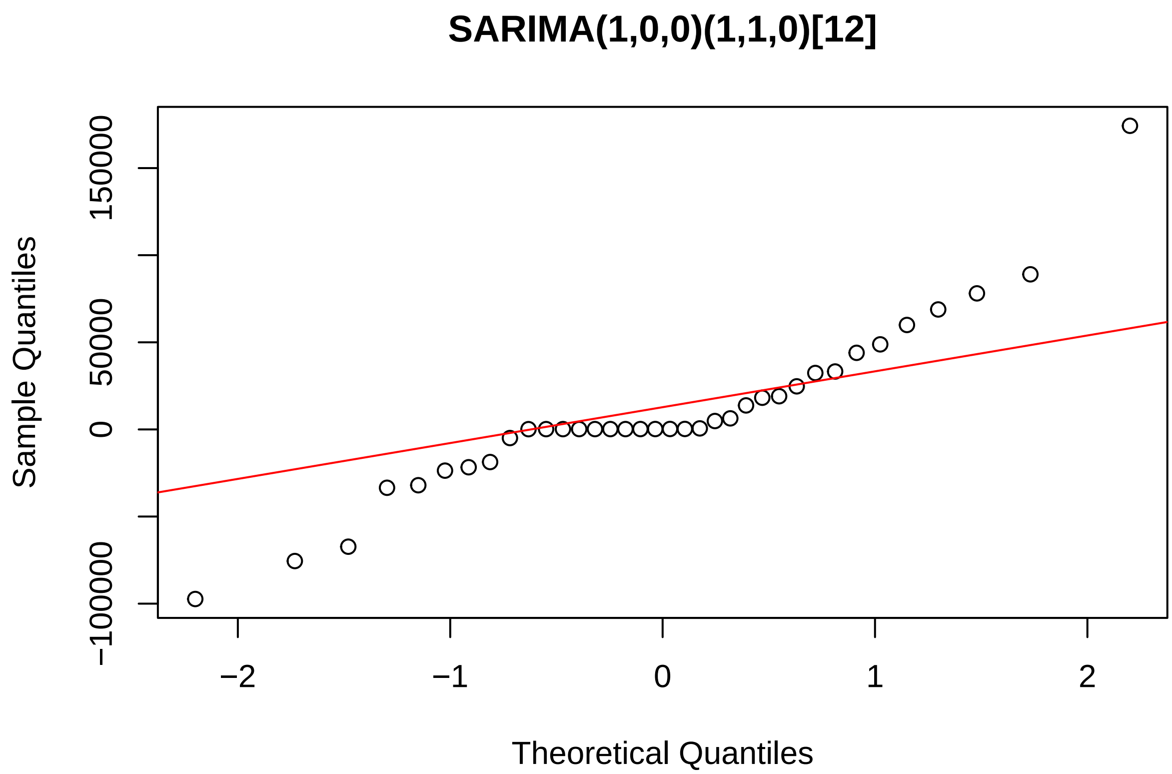
De la misma forma, para el modelo SARIMA(1,0,1)(0,1,0)[12]




```
##
##  Shapiro-Wilk normality test
##
## data:  residuos2
## W = 0.88559, p-value = 0.001409
```

tampoco cumple el supuesto de normalidad para los residuos.

Finalmente para el modelo SARIMA(1,0,0)(1,1,0)[12]




```
##
##  Shapiro-Wilk normality test
##
## data:  residuos3
## W = 0.90396, p-value = 0.004385
```

vemos que tampoco cumple el supuesto de normalidad.

Test de blancura

Para complementar esto, realizaremos un test de Box-Ljung en donde:

- H_0: Los residuos son independientes
- H_1: Los residuos no son independientes

Rechazamos H_0 si el p-value es menor a nuestro α del 5%.

- 1.- SARIMA(1,0,0)(0,1,0)[12]

```
##
##  Box-Ljung test
##
## data:  residuos1
## X-squared = 2.2437, df = 1, p-value = 0.1342
```

Tenemos que el p-value es mayor a nuestro α, por lo que tenemos evidencia suficiente para no rechazar H_0. Por lo tanto, los residuos son independientes y corresponden a un ruido blanco.

- 2.- SARIMA(1,0,1)(0,1,0)[12]

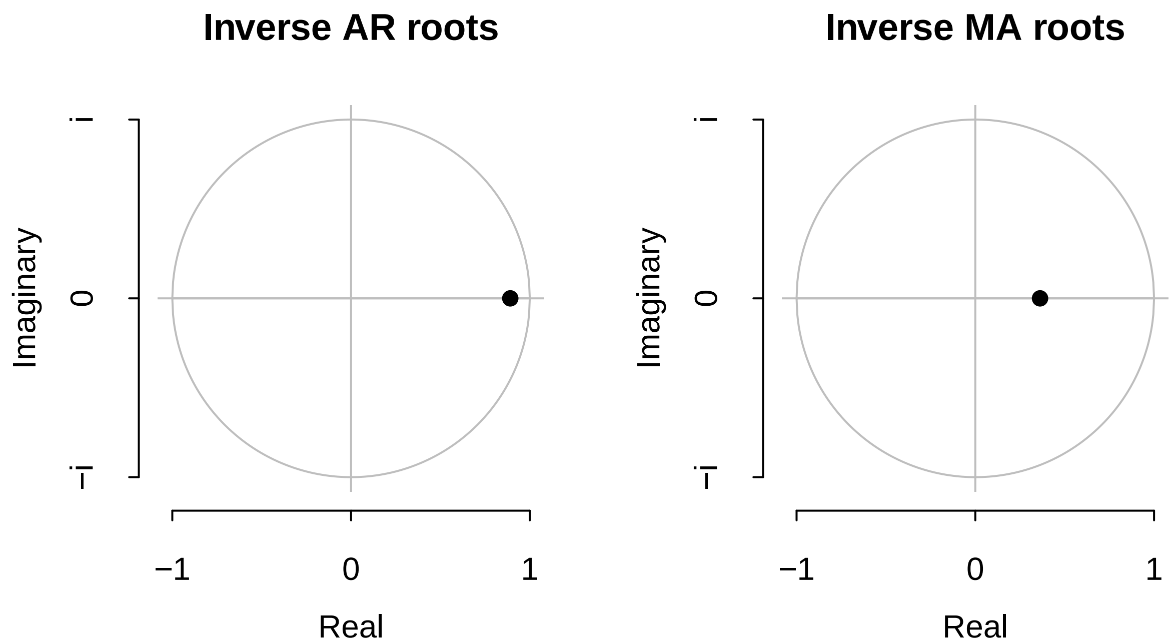
```
##
##  Box-Ljung test
##
## data:  residuos2
## X-squared = 0.00025175, df = 1, p-value = 0.9873
```

- 3.- SARIMA(1,0,0)(1,1,0)[12]

```
##
##  Box-Ljung test
##
## data:  residuos3
## X-squared = 1.3715, df = 1, p-value = 0.2416
```

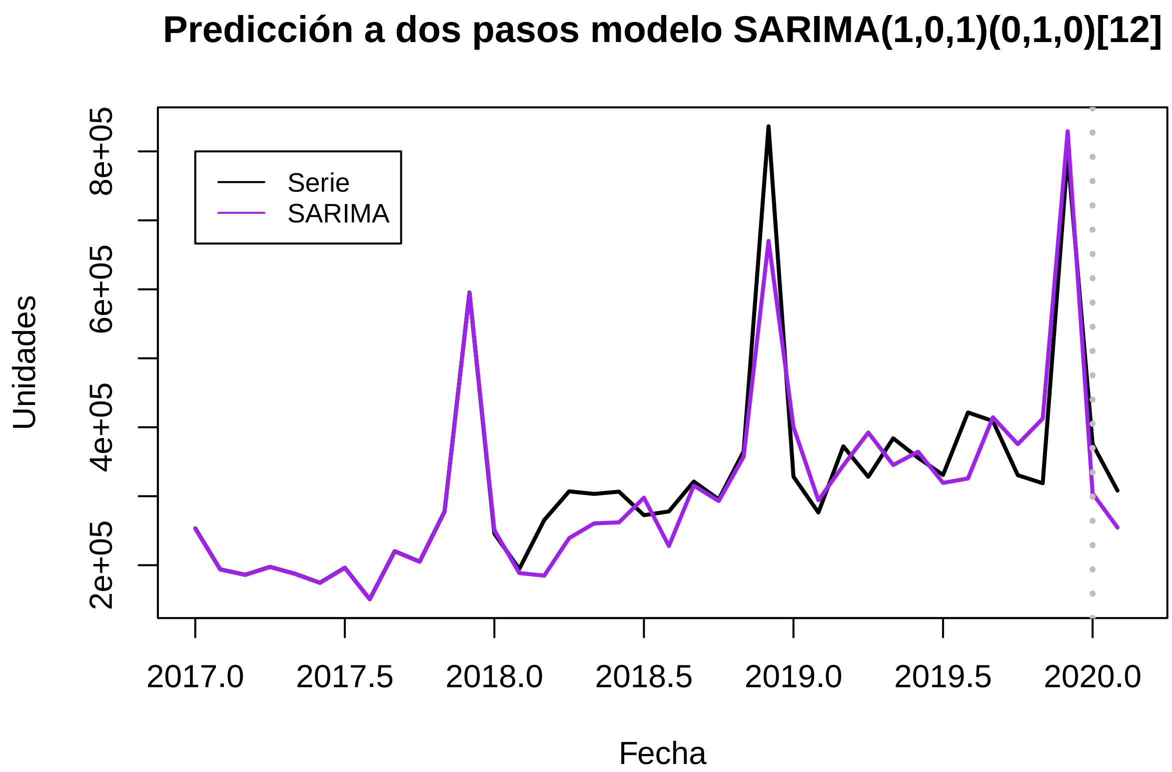
Tenemos que el p-value es mayor a nuestro α, por lo que tenemos evidencia suficiente para no rechazar H_0. Por lo tanto, los residuos son independientes y corresponden a un ruido blanco para los tres modelos. Sin embargo, el modelo que mejor cumple este supuesto es el modelo **SARIMA(1,0,1)(0,1,0)[12]**, por lo cual nos quedamos con este modelo para analizar.

Modelo Causal e Invertible



Se puede apreciar que las raíces se encuentran dentro del círculo unitario, es decir, el modelo es causal e invertible y por lo tanto se pueden hacer predicciones.

6.- Predicciones a dos pasos:

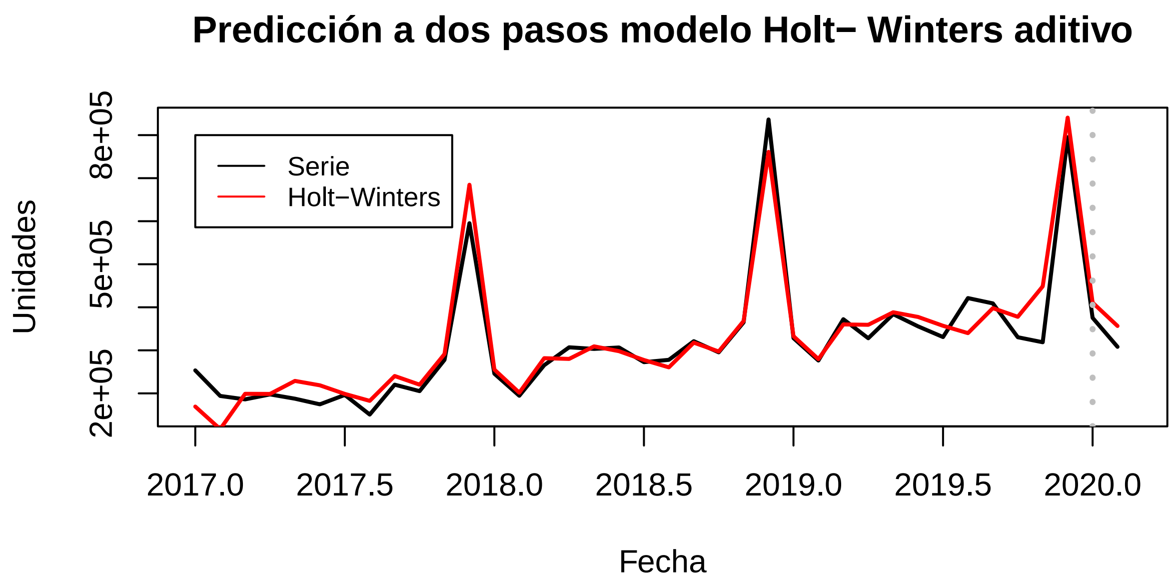


Donde los valores predichos para el mes de enero y febrero son 303979.0 y 254605.9 respectivamente.

7.- Modelo ingenuo

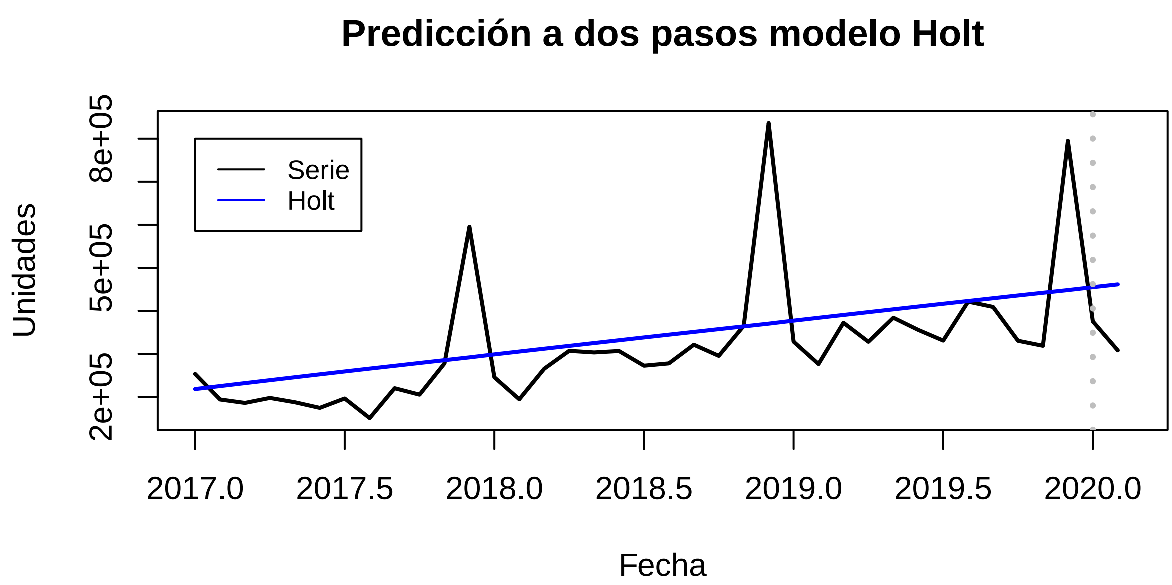
Se ajustaron tres modelos ingenuos los cuales son los siguientes

- 1.- Holt-Winters



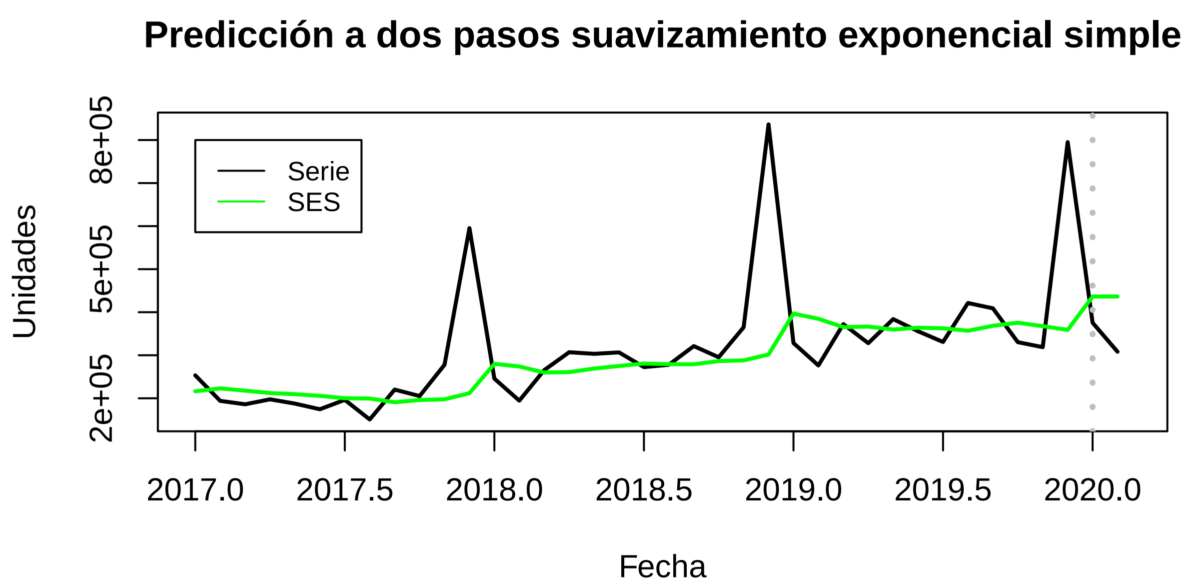
donde sus parámetros ajustados fueron $\alpha = 0.0048$, $\beta = 0.0048$ y $\gamma =$ 1e-04. Los valores predichos para el mes de enero y febrero son 410847.0 y 356603.2 respectivamente.

- 2.- Holt



donde sus parámetros ajustados fueron α = 7e-04 y β = 7e-04. Los valores predichos para el mes de enero y febrero son 454899.0 y 461402.8 respectivamente.

- 3.- Suavizamiento exponencial simple (SES)



donde su parámetro ajustado fue α = 0.1778 y los valores predichos para el mes de enero y febrero son 436482.6 y 436482.6 respectivamente.

En base a lo anterior, seleccionamos el modelo ingenuo Holt-Winters, ya que capta de mejor manera la estacionalidad de la serie.

8.- Comparación de modelos

Para comparar los modelos, utilizaremos la métrica del MAPE para comparar el puntaje de cada modelo.

```
##
## Forecast method: Holt-Winters' additive method
##
## Model Information:
## Holt-Winters' additive method
##
## Call:
##  hw(y = dt, h = 2)
##
##   Smoothing parameters:
##     alpha = 0.0048
##     beta  = 0.0048
##     gamma = 1e-04
##
##   Initial states:
##     l = 197749.5264
##     b = 6947.2172
##     s = 400037.9 13442.79 -50883.15 -23814.81 -74651.58 -51274.57
##            -24311.41 -6791.582 -29404.48 -21556.8 -95424.89 -35367.43
##
##   sigma:  56645.14
##
##      AIC     AICc      BIC
## 929.8548 963.8548 956.7746
##
## Error measures:
##                     ME    RMSE      MAE       MPE     MAPE     MASE      ACF1
## Training set -6605.465 42220.8 28675.54 -2.627527 9.629199 0.343791 0.2952582
##
## Forecasts:
##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Jan 2020       410847.0 338253.3 483440.6 299824.5 521869.4
## Feb 2020       356603.2 284006.3 429200.2 245575.8 467630.7
```

Se muestran los parámetros ajustados para el modelo HW y el valor de cada métrica.

```
##
## Forecast method: ARIMA(1,0,1)(0,1,0)[12]
##
## Model Information:
## Series: dt
## ARIMA(1,0,1)(0,1,0)[12]
##
## Coefficients:
##          ar1      ma1
##       0.8908  -0.3621
## s.e.  0.0983   0.2597
##
## sigma^2 = 3.595e+09:  log likelihood = -297.51
```

```
## AIC=601.03   AICc=602.23   BIC=604.56
##
## Error measures:
##                    ME     RMSE      MAE      MPE     MAPE      MASE
## Training set 7739.632 46870.58 28379.44 1.699851 7.834837 0.3402411
##                      ACF1
## Training set -0.002537894
##
## Forecasts:
##          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## Jan 2020       303979.0 227140.9 380817.0 186465.3 421492.6
## Feb 2020       254605.9 167690.1 341521.8 121679.6 387532.3
```

Se muestran los parámetros ajustados para el modelo SARIMA y el valor de cada métrica.

En vista de lo anterior, el mejor modelo según el MAPE es el modelo SARIMA

9.- Conclusiones

En base a lo anterior, podemos ver que el ajuste del modelo SARIMA es el que tiene un menor MAPE en el ajuste de la serie, por ende las predicciones a dos pasos se acercan mejor a la venta real que se tendría para los meses de enero y febrero del año 2020. Lo anterior, ayuda en conocer como afecta el mes en que nos encontramos con la venta generada en unidades, lo cual es útil para la empresa en términos de lógistica, para saber como redistribuir los productos vendidos anualmente.