

0.1. Modelos estadísticos clásicos

0.1.1. Modelo de Regresión Lineal General

0.1.2. Análisis de Series Temporales

Se consideró aplicar un análisis de series temporales debido a la estructura de los datos, ya que este tipo de análisis contempla la estructura temporal de los mismos. Como ya se avanzó en el desarrollo teórico, aplicaremos la metodología Box-Jenkins, la cual tiene en cuenta la dependencia existente de los datos, construyendo así un modelo ARIMA.

Trataremos de modelizar el volumen de ventas total según día de la semana. Para construir la serie, primero hemos añadido los días 25 de Diciembre y 1 de Enero con un número de ventas 0, ya que, si no se tomaba esta decisión, la serie ya no estaría definida según la realidad.

0.1.2.1. Creación ST y representación de los datos

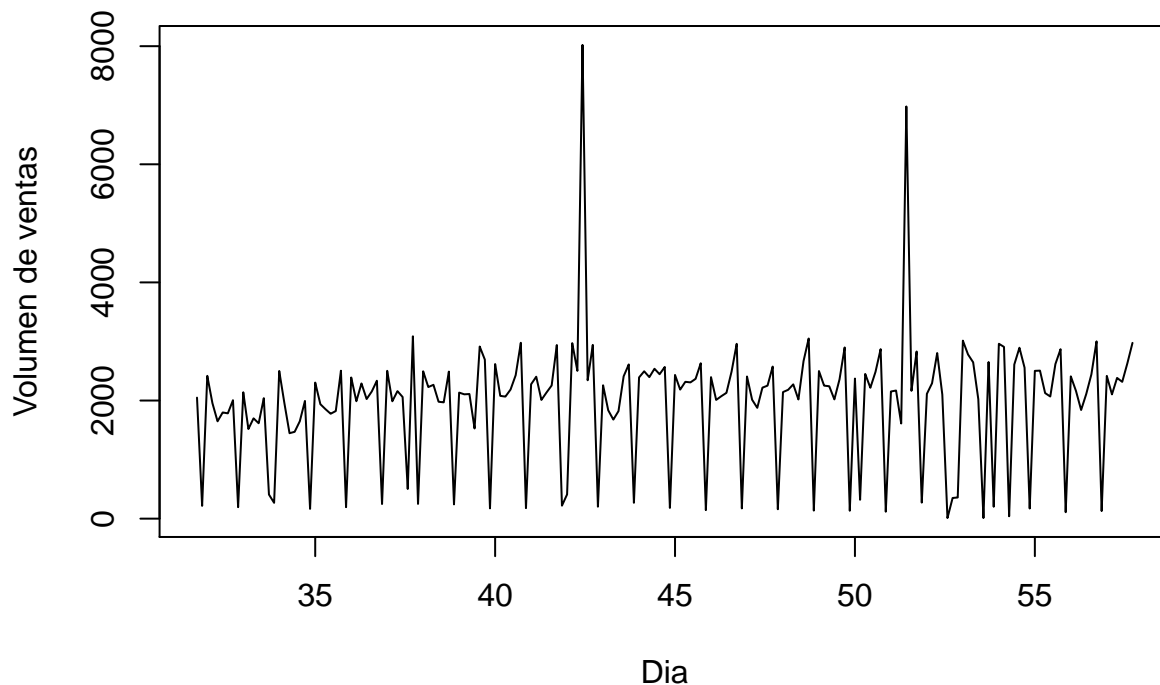
Si construimos la serie con los valores actuales, no podremos aplicar transformaciones a la serie, en particular la transformación de Box-Cox, ya que existen dos valores nulos. Por este motivo, sumamos una constante a todas las observaciones de modo que sean todas positivas.

##	p1	p2	p3	p4	p5	p6	p7
## 31						2049	216
## 32	2418	1955	1648	1798	1783	2005	194
## 33	2140	1518	1698	1615	2040	408	268
## 34	2501	1957	1447	1470	1649	1993	165
## 35	2304	1938	1853	1776	1822	2505	194
## 36	2390	1988	2289	2025	2153	2335	247
## 37	2503	1987	2159	2060	502	3088	249
## 38	2495	2228	2266	1979	1969	2490	241
## 39	2134	2105	2110	1527	2914	2693	174
## 40	2618	2079	2066	2186	2429	2978	178
## 41	2275	2403	2010	2136	2255	2940	220
## 42	410	2972	2502	8021	2343	2942	202
## 43	2258	1835	1677	1821	2410	2611	268
## 44	2391	2495	2396	2538	2445	2568	181
## 45	2433	2185	2317	2307	2367	2631	144
## 46	2396	2010	2073	2130	2491	2960	173
## 47	2407	2013	1876	2218	2252	2575	159
## 48	2143	2181	2274	2018	2667	3050	136
## 49	2499	2252	2242	2019	2358	2899	133
## 50	2372	319	2450	2216	2486	2870	117
## 51	2153	2169	1611	6979	2164	2830	271
## 52	2113	2292	2804	2100	10	350	358
## 53	3014	2780	2647	2023	10	2652	201
## 54	2961	2908	40	2609	2894	2556	171
## 55	2502	2506	2126	2068	2617	2869	109
## 56	2409	2163	1840	2110	2443	3002	128

```
## 57 2417 2104 2383 2316 2631 2976
```

Después de haber definido los datos como una serie temporal, visualizamos la evolución de la serie en el tiempo.

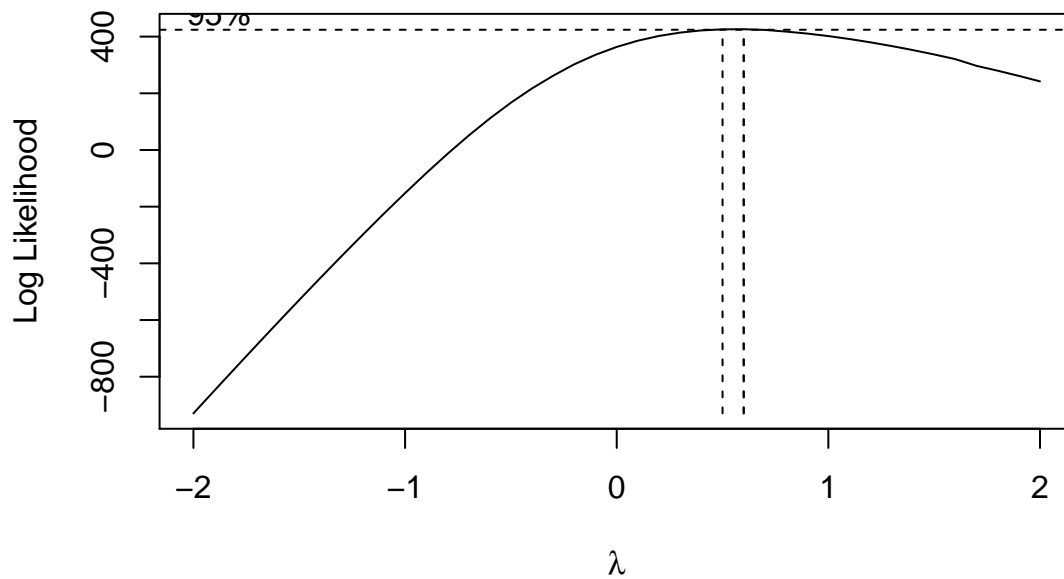
Volumen total de venta por según día de la semana



En el gráfico se puede apreciar cierta estacionalidad de los datos, es decir, movimientos que se repiten regularmente año tras año en los mismo períodos. También observamos que las oscilaciones van aumentando con el tiempo, indicando que la varianza no es constante. Por este motivo, debemos hacer alguna transformación para que la varianza sea constante en el tiempo.

0.1.2.2. Transformación de BoxCox para estabilizar la varianza

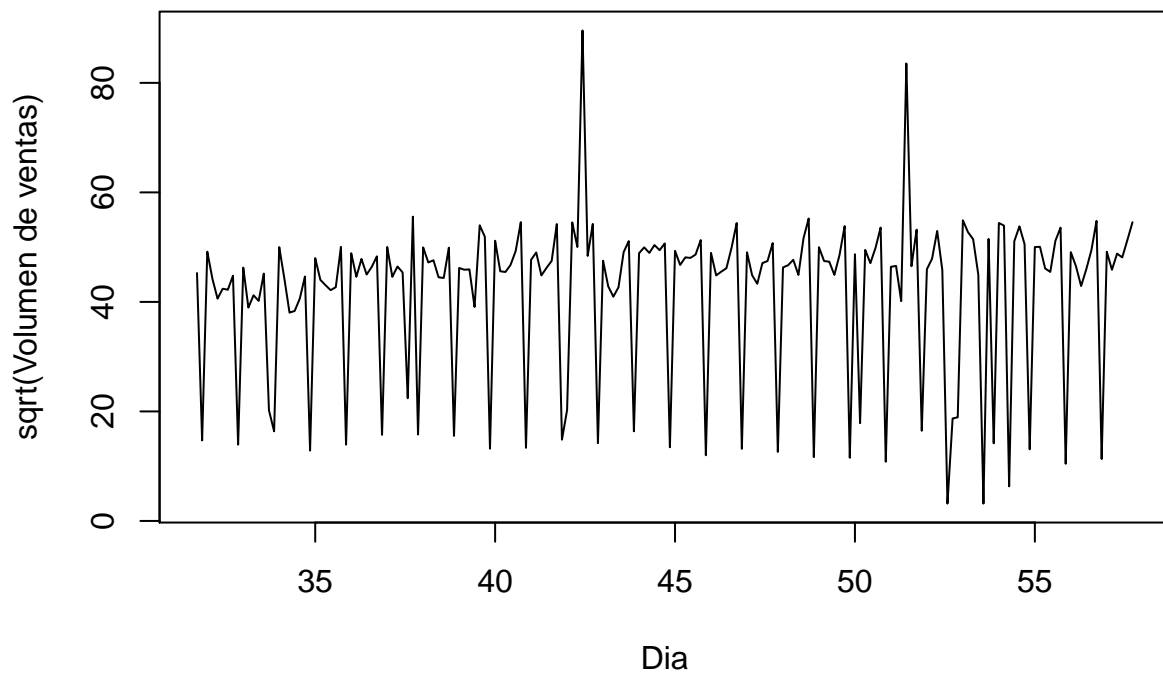
Para encontrar una transformación que haga que la varianza sea constante en el tiempo, haremos uso de la familia de transformaciones Box-Cox con ayuda de la librería *TSA*.



La función *BoxCox.ar* sugiere un óptimo de $\lambda = 0.6$, con un intervalo de confianza al 95 %: (0.5,0.6). Se necesita una transformación sencilla y comprensible, por lo que se ha optado por tomar como valor de lambda el extremo inferior del intervalo, $\lambda = 1/2$.

Transformamos los datos y volvemos a representar la serie.

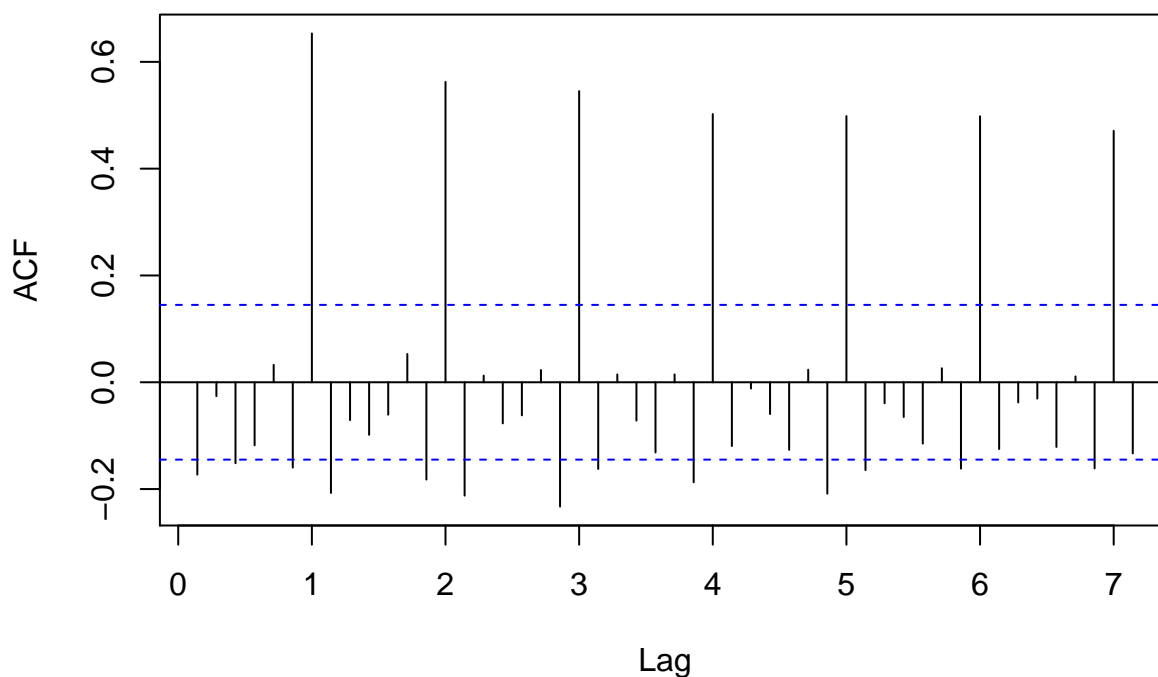
sqrt(Volumen total de venta diario)



0.1.2.3. Transformaciones para estabilizar la media

Vamos a estudiar si el motivo de la no estacionalidad de los datos en media se debe a que se trata de un proceso integrado. Para ello, hacemos uso de la función de autocorrelación simple.

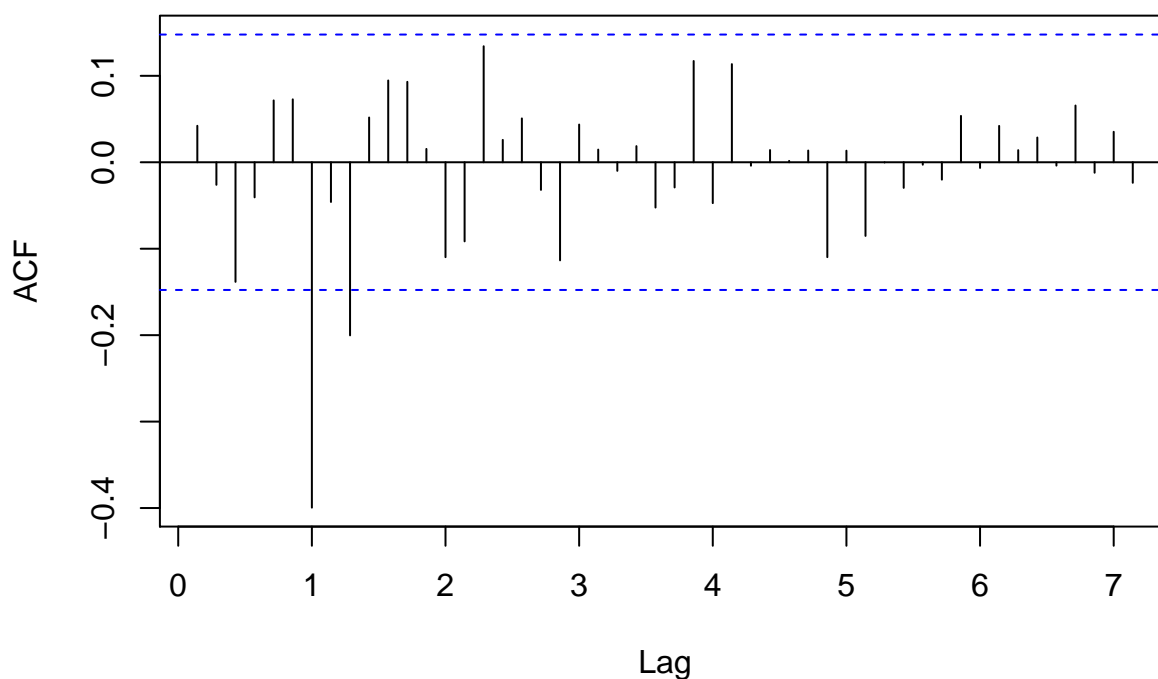
FAS de SQRT de Ventas diarias



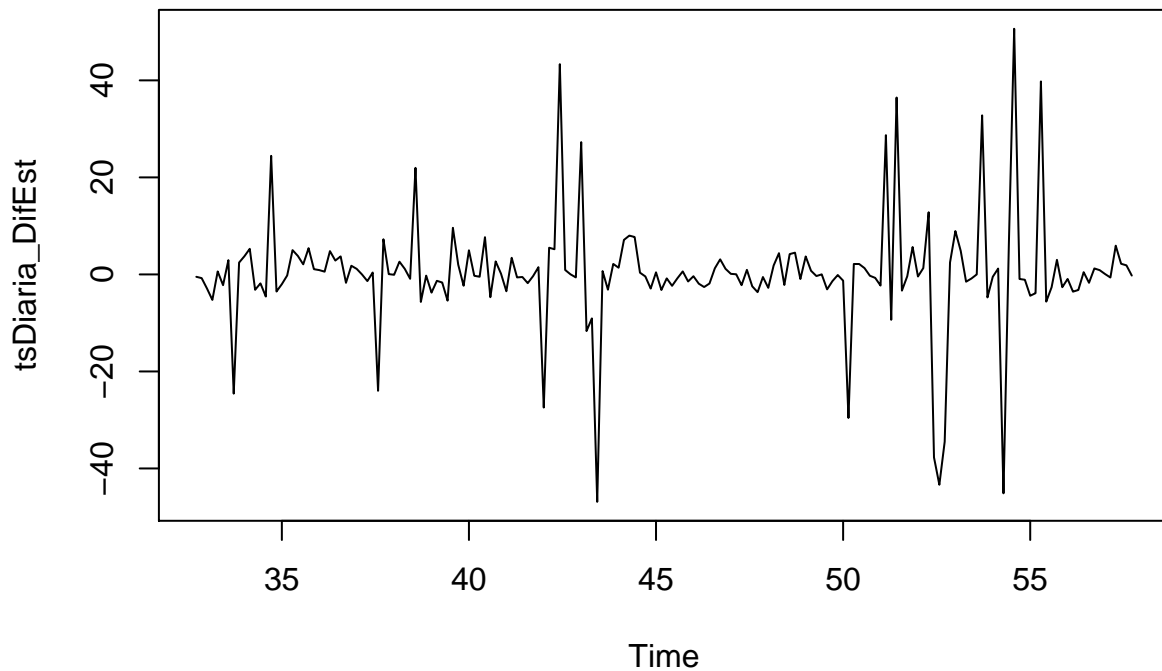
La FAS muestral decrece de lentamente en los retardos estacionales de período 7, indicando que estamos ante un modelo integrado. Debido a esta situación, hacemos una diferencia estacional de la serie y volvemos a representar la FAS ($s=7$).

```
tsDiaria_DifEst = diff(tsDiaSemanal_transf, lag=7, diff=1)
acf(tsDiaria_DifEst, main="FAS de primera diferencia estacional", lag=50)
```

FAS de primera diferencia estacional



Ahora la función de autocorrelación muestral corresponde a la de un proceso estacionario. Por último, representamos gráficamente la serie diferenciada:



Observamos que la serie no muestra ningún comportamiento en particular, sino que se aprecia aleatoriedad, por lo que se podría pensar, que nos encontramos ante un proceso estacionario. Ahora estamos en condiciones de buscar un modelo estacionario para la serie.

0.1.2.4. Contraste de estacionariedad

Para confirmar la estacionariedad de los datos sugerida con la observación de la gráfica, necesitamos aplicar un test de hipótesis. Aplicamos el test de raíz unitaria de Dikey-Fuller, donde se contrasta la estacionariedad de los datos a través del siguiente test de hipótesis:

$$\begin{cases} H_0 : & \text{El polinomio autoregresivo tiene una raíz unitaria} \\ H_1 : & \text{Todas las raíces del polinomio autoregresivo son estacionarias} \end{cases}$$

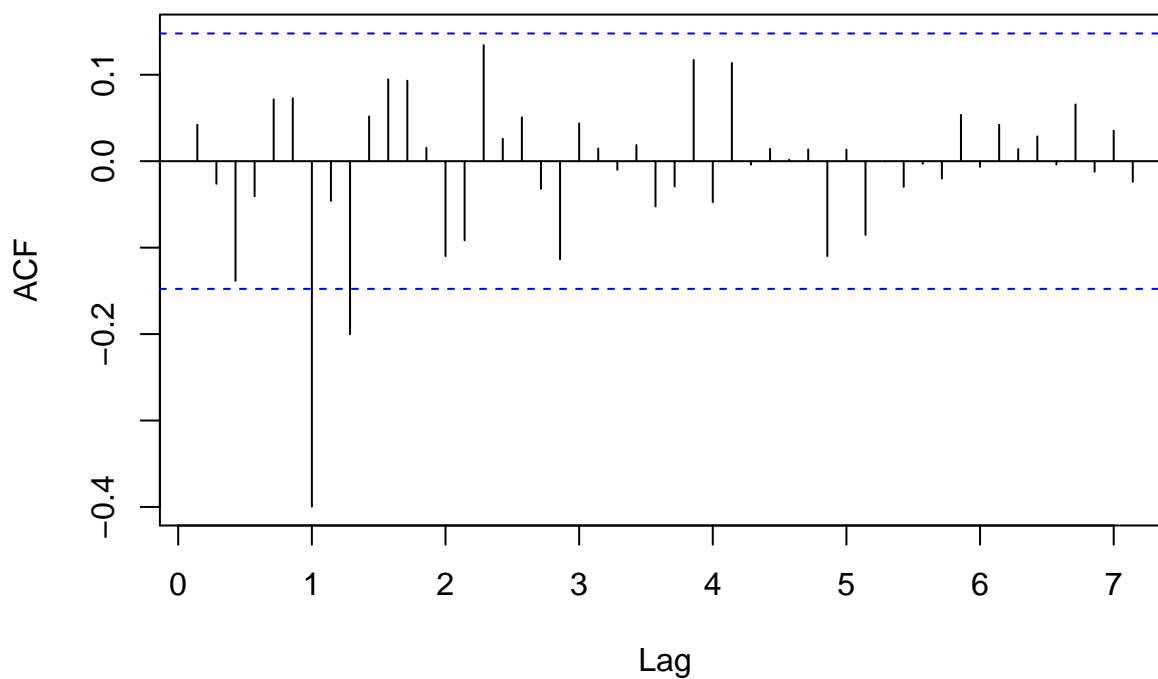
```
##
## Augmented Dickey-Fuller Test
##
## data: tsDiaria_DifEst
## Dickey-Fuller = -5.1008, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

El p-valor del test= 0.01 < 0.05 = α , y por tanto concluimos que no existen evidencias significativas para asumir que el polinomio autoregresivo tiene alguna raíz unitaria, la serie es estacionaria.

0.1.2.5. Identificación de la estructura ARIMA de la serie

Trataremos de identificar la estructura ARIMA más adecuada para esta serie a través de la función de autocorrelación simple (FAC) y de la función de autocorrelación parcial (FAP). Determinar el modelo más adecuado consistirá en e identificar el orden de los procesos de medias móviles y autoregresivos de la componente estacional y la componente regular.

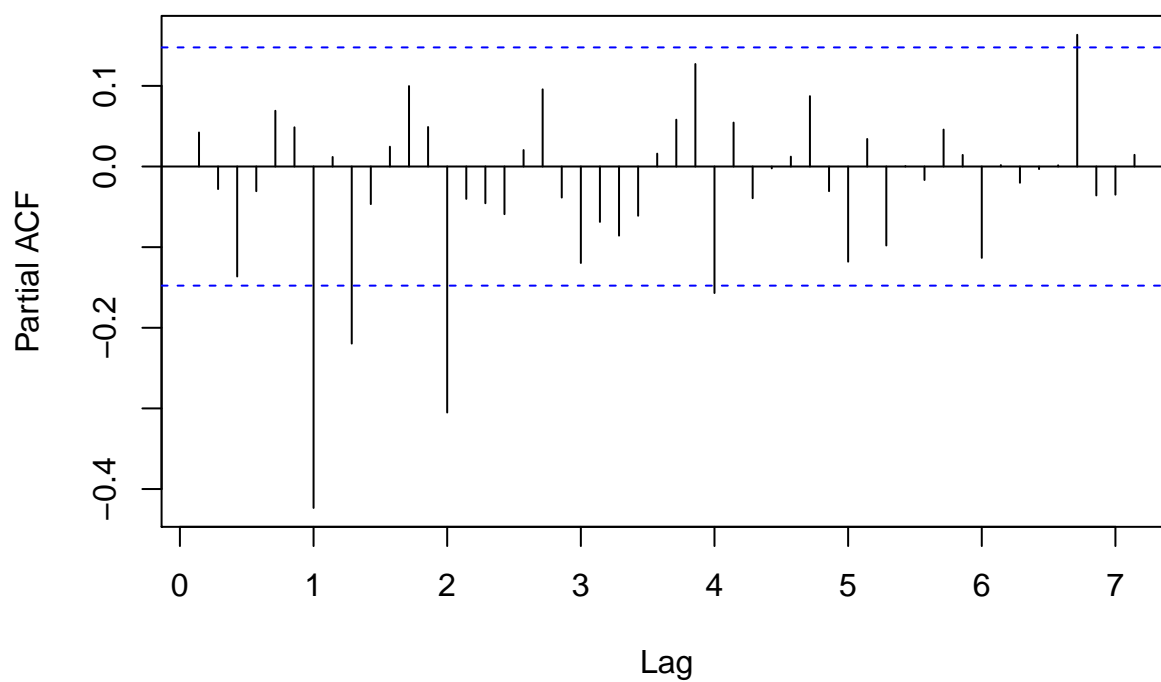
FAS tras una diferencia estacional



- Parte regular: En los primeros retardos no observamos ninguna autocorrelación significativamente no nula, indicando que el modelo tiene una estructura $ARMA(0,0)$ en la parte regular.
- Parte estacional: Observamos una autocorrelación en el primer retardo estacional, por lo que parecería que la parte estacional tiene una estructura $MA(1)_{12}$.

Vamos a comprobar estos supuestos con la FAP.

FAP tras una diferencia estacional



- Parte regular: De nuevo, no hay autocorrelaciones significativamente no nulas en los primeros retardos.
- Parte estacional: En los retardos estacionales, observamos como las autocorrelaciones decrecen rápidamente y a su izquierda, no hay autocorrelaciones significativamente no nulas, lo que avalaría aún más la suposición de un $MA(1)$ en la parte estacional. Modelo propuesto: $MA(1)_{12}$

También observamos como hay otras autocorrelaciones significativamente no nulas, pero esto es debido a que se trata de un intervalo de confianza al 95 %, por lo que cabe esperar que haya algunas autocorrelaciones fuera de las bandas.

El modelo a considerar es un modelo estacional multiplicativo integrado de medias móviles puro: $ARIMA(0, 1, 1)_{12}$

0.1.2.6. Estimación de parámetros y diagnóstico del modelo

Una vez hemos obtenido un modelo, se han estimado sus parámetros con la función *arima*.

```
Ajuste1 = arima( tsDiaria_DifEst ,    # Serie tras una diferencia estacional
                seasonal = list(order=c(0,0,1),period=7 ))
Ajuste1

##
## Call:
## arima(x = tsDiaria_DifEst, seasonal = list(order = c(0, 0, 1), period = 7))
##
## Coefficients:
##          sma1  intercept
##         -1.0000      0.0898
## s.e.      0.0827      0.0871
##
## sigma^2 estimated as 79.04:  log likelihood = -645.71,  aic = 1295.42

Trás comprobar si los coeficientes estimados son o no significativamente nulos, procedemos
a eliminar la media del modelo, obteniendo así uno donde todos los coeficientes son
significativamente no nulos.

##              2.5 %      97.5 %
## sma1         -1.16213365 -0.8378657
## intercept    -0.08097887  0.2605421

Ajuste1_1 = arima( tsDiaria_DifEst ,    # Serie tras una diferencia estacional
                  order = c(0,0,0),seasonal = list(order=c(0,0,1),period=7),
                  include.mean = FALSE # Eliminamos la media
                  )
Ajuste1_1

##
## Call:
## arima(x = tsDiaria_DifEst, order = c(0, 0, 0), seasonal = list(order = c(0,
##      0, 1), period = 7), include.mean = FALSE)
##
```

```
## Coefficients:
##          sma1
##          -0.9999
## s.e.      0.1009
##
## sigma^2 estimated as 79.52:  log likelihood = -646.24,  aic = 1294.47
confint(Ajuste1_1)

##          2.5 %      97.5 %
## sma1 -1.197648 -0.8021465
```

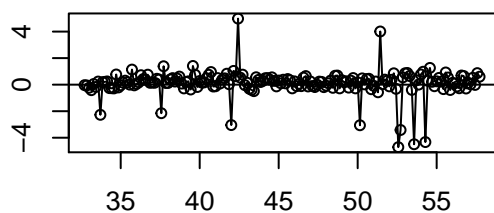
El modelo ajustado corresponde a la siguiente ecuación:

$$Y_t = (1 - L^7)(1 + 0.9999\Theta^7)\alpha_t, \quad \alpha_t \sim RB(0, \sigma^2)$$

Para comprobar si el modelo es o no adecuado, comprobamos su validez a través de la diagnosis de los residuos y concluimos que este ajuste no es adecuado, ya que según el Test de Ljung-Box, no existen evidencias significativas para aceptar la incorrelación de los residuos: $p\text{-valor} = 0.002524 < 0.05 = \alpha$. Además, gráficamente podemos observar que los residuos no se comportan como un ruido blanco.

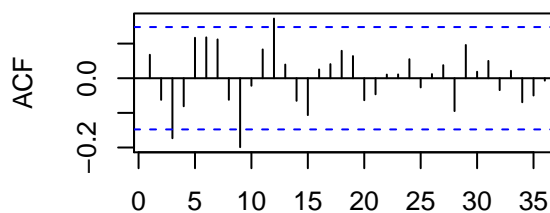
```
##
## Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,0)(0,0,1)[7] with zero mean
## Q* = 31.855, df = 13, p-value = 0.002524
##
## Model df: 1.   Total lags used: 14
```

Residuos estandarizados



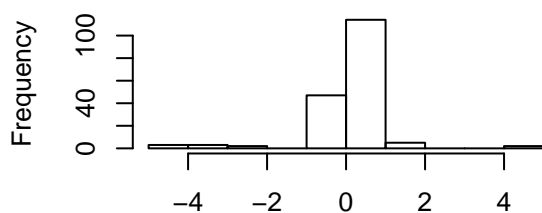
Time

FAS de los residuos



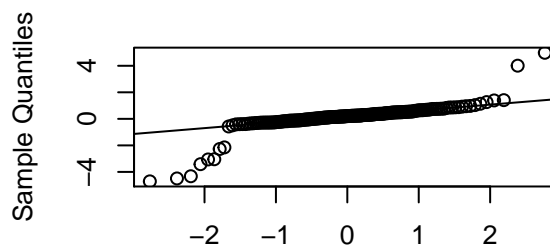
Lag

Histograma



Residuos estandarizados

Normal Q-Q Plot

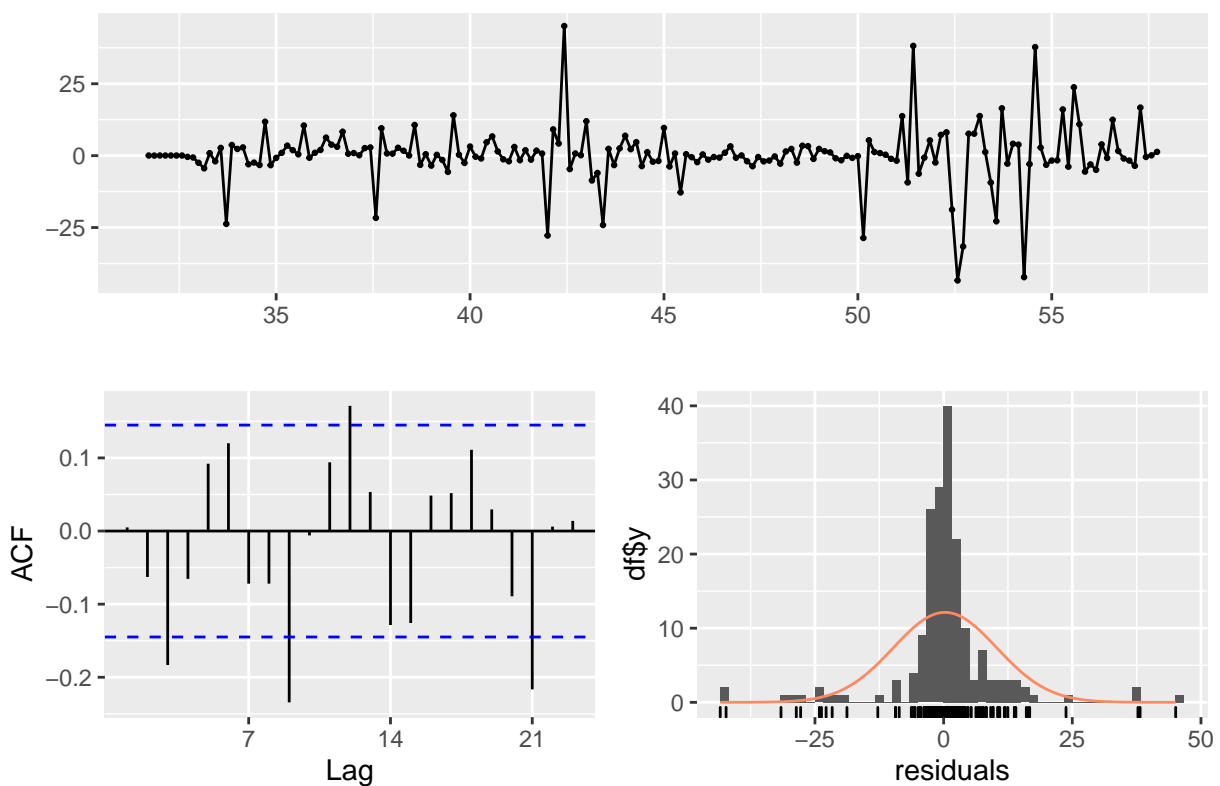


Theoretical Quantiles

Vamos a probar otro modelo, en particular, a través del paquete *forecast* haciendo uso de la función *auto.arima*, que busca un modelo que minimiza el AIC.

```
## Series: tsDiaSemanal_transf
## ARIMA(1,0,0)(2,1,0)[7]
##
## Coefficients:
##          ar1      sar1      sar2
##          0.0709  -0.5180  -0.3068
## s.e.    0.0754   0.0708   0.0694
##
## sigma^2 = 106.2:  log likelihood = -660.08
## AIC=1328.17  AICc=1328.4  BIC=1340.85
```

Residuals from ARIMA(1,0,0)(2,1,0)[7]



```
##
## Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0)(2,1,0)[7]
## Q* = 36.336, df = 11, p-value = 0.0001488
##
## Model df: 3.    Total lags used: 14
```

El ajuste propuesto es un modelo: $ARIMA(1, 0, 0) \times ARIMA(2, 0, 0)_7$, pero tampoco es adecuado, ya que volvemos a rechazar la hipótesis de incorrelación de los residuos del Test de Ljung-Box.

No hemos podido encontrar un modelo adecuado que se ajuste a los datos y que pase la diagnosis, ya que los residuos no provenían en ningún caso de un proceso de ruido blanco, es decir, no estaban incorrelados entre sí. Por este motivo, al no ser los retardos

	MODELO	AIC	p-valor
Modelo 1	ARIMA(0,1,1)_7	1294.474	0.0025242
Modelo 2	ARIMA(1,0,0)x(2,1,0)_7	1328.403	0.0001488
Modelo 3	ARIMA(1,1,1)_7	1293.397	0.0068141
Modelo 4	ARIMA(1,1,0)_7	1341.251	0.0000002

Note:

El p-valor corresponde al test de Ljung-Box

independientes, un retardo puede guardar cierta relación con otro retardo k períodos después. En estos casos, la autocorrelación puede conducir a una inexactitud en el modelo predictivo, que nos llevaría a interpretaciones erróneas.

La tabla mostrada a continuación expone los diferentes modelos ajustado, el valor del AIC y el p-valor obtenido del test de Ljung-Box. De haber pasado algún modelo la diagnosis, el seleccionado para realizar predicciones del volumen de ventas habría sido aquel con menor valor del AIC.