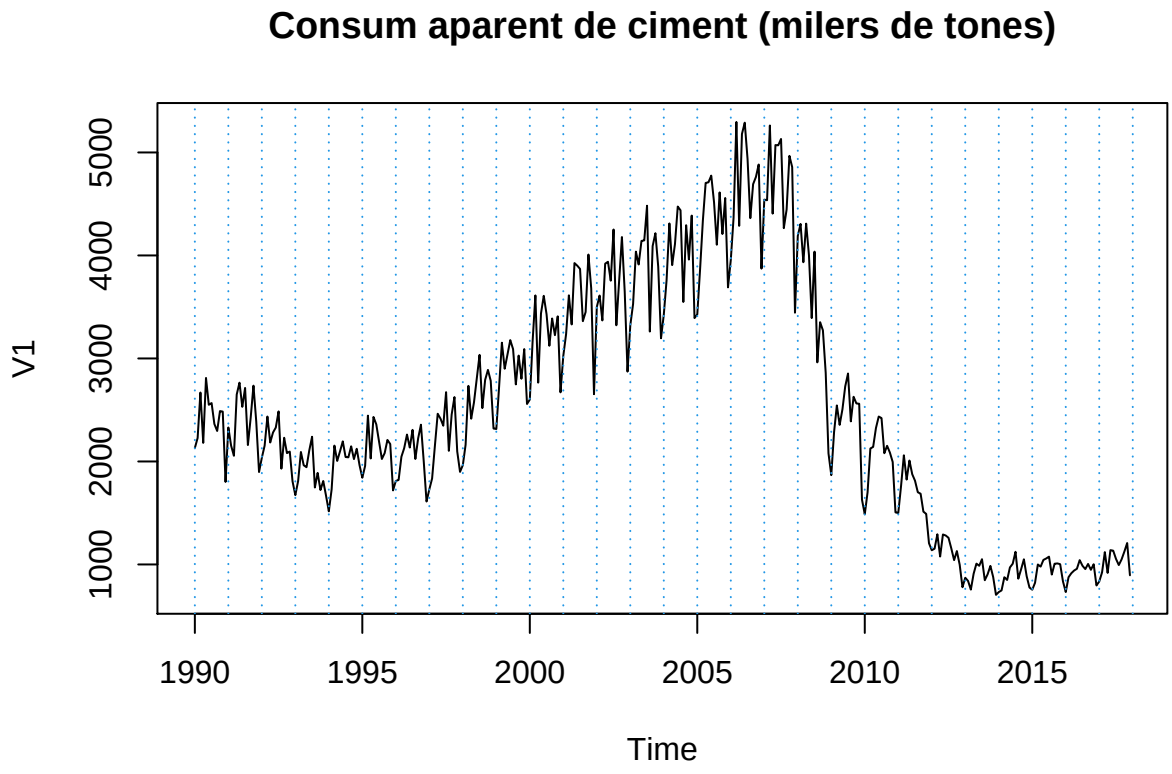
Treball Sèries Temporals

Sergio Cárdenas & Armand de Asís

25/4/2022

Preparació

Començarem analitzant la sèrie temporal "cemento.dat". Veiem que són dades mensuals sobre el consum de ciment a Espanya des de 1990 fins a 2018. Apliquem una correcció sobre les dades perque mostrin milers de tones per així evitar que a l'hora d'analitzar la variància els valors siguin excesivament grans.



Estadística descriptiva de la sèrie

Fins als darrers anys del segle XX sembla haver un consum de ciment relativament semblant, que als anys propers al 2000 conmença a créixer fins al 2008, on decreix radicalmen. Al 2008, podem veure clarament l'efecte de la crisi amb una davallada del consum de ciment. Possiblement serà un outlier que haurem de

considerar. Cap al 2013 sembla que el consum de ciment es queda constant i a valors molt petits, degut a que la construcció de nous habitatges es va reduir després de que esclatés la bombolla immobiliària.

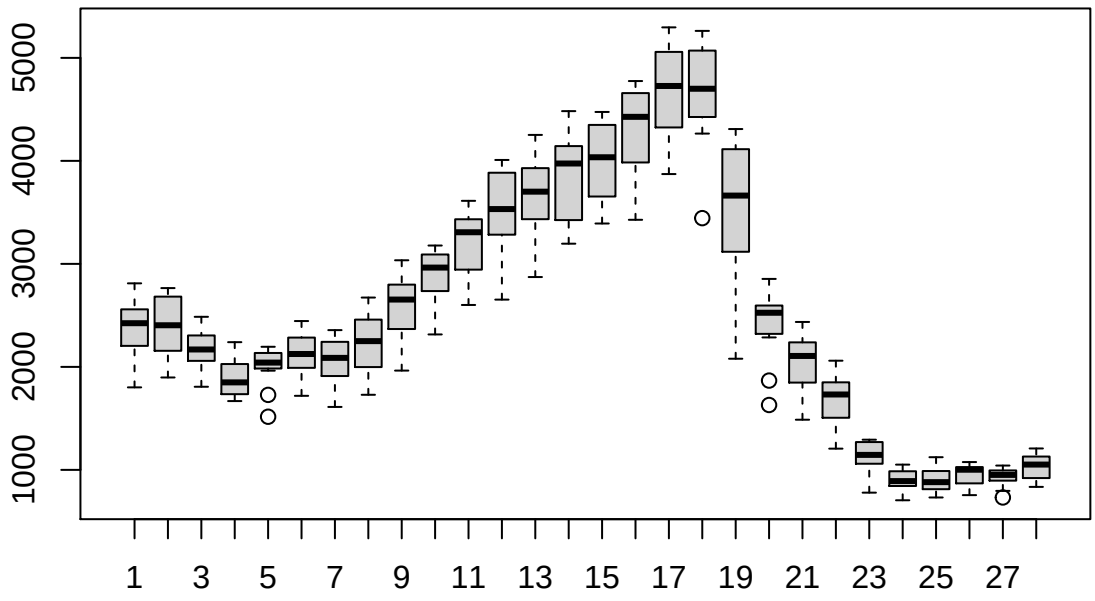
A més a més, podem veure que hi ha un cert comportament estacional, amb un creixement en el consum de ciment a les èpoques estivals, i no sembla haver variància constant al llarg de la sèrie.

Estacionarietat de la sèrie

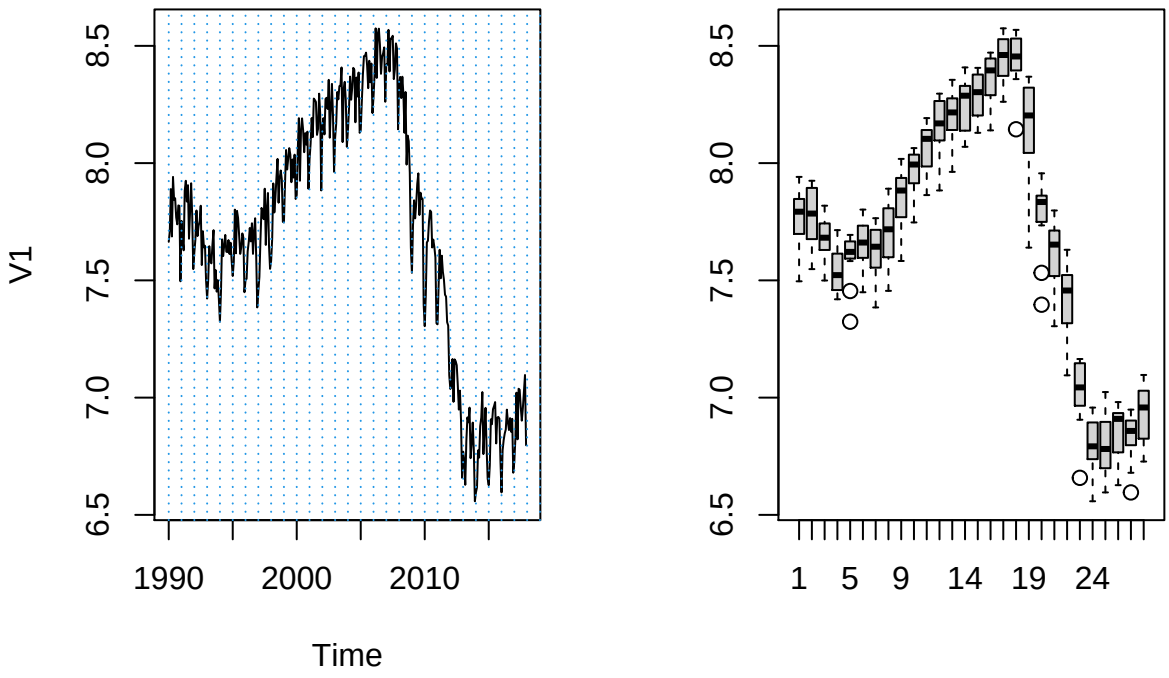
Estudi de l'homocedasticitat

Farem ara les transformacions necessàries per aconseguir l'estacionarietat de la sèrie.

Fent un boxplot de la sèrie (on cada capsa representa un any amb 12 observacions pels 12 messos), es pot veure que les variàncies no són constants, ja que les "capses" que representen valors més alts són molt més amples que les dels valors petits, per tant, haurem d'aplicar una transformació Box-Cox, en aquest cas el logaritme.

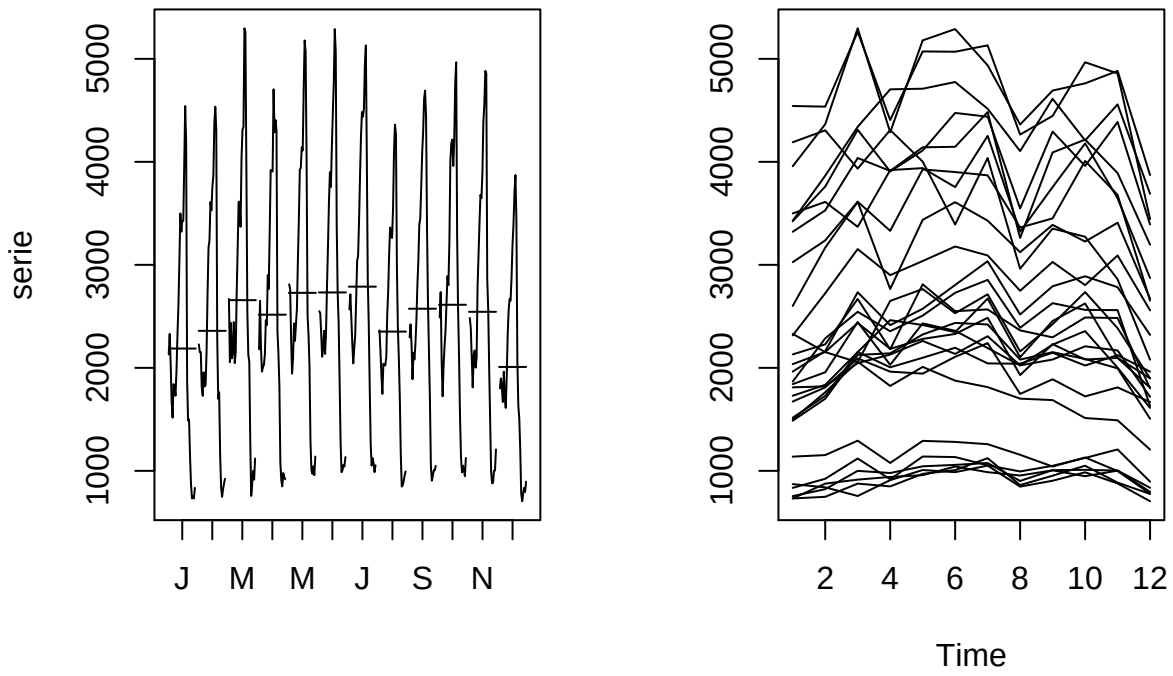


Apliquem la transformació logarítmica per tal d'estabilitzar les variàncies i podem veure que ha funcionat:

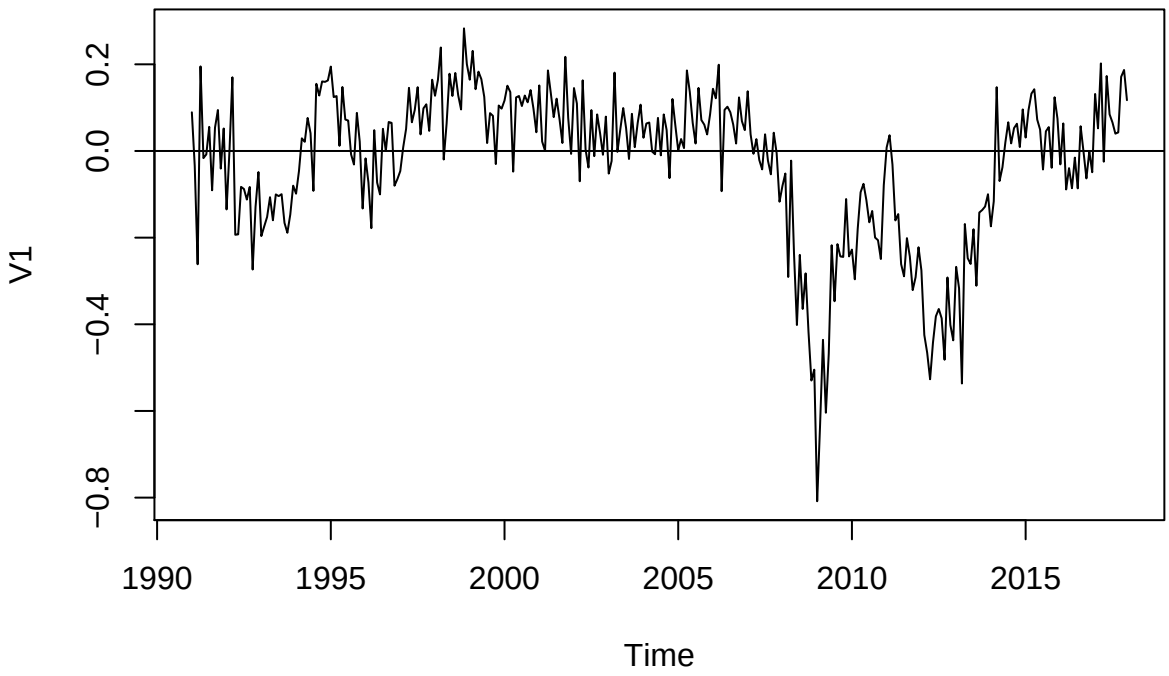


Estudi de l'estacionalitat

Com hem vist a l'estadística descriptiva, la sèrie sembla presentar un comportament estacional. Per tant, comprovem que és així:

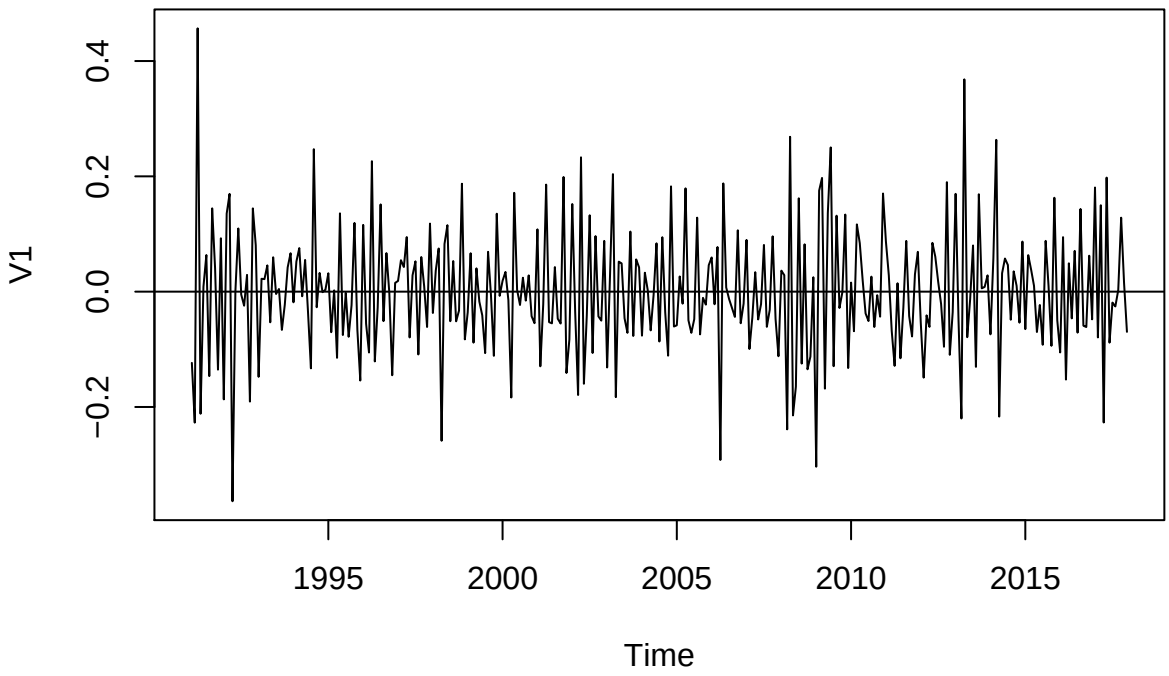


Podem veure que, tot i que no sembla molt evident a tots els casos, hi ha un patró estacional que es caracteritza per valors lleugerament més petits a l'hivern, probablement degut a que les condicions climàtiques dificulten el treball a l'aire lliure, de manera que és evident que amb el bon temps vingui l'augment en el consum de ciment, a excepció d'agost, més al qual s'aturen els treballs per vacances. Per tant, hem d'aplicar diferenciació estacional de període 12 amb l'objectiu de corregir aquest comportament:

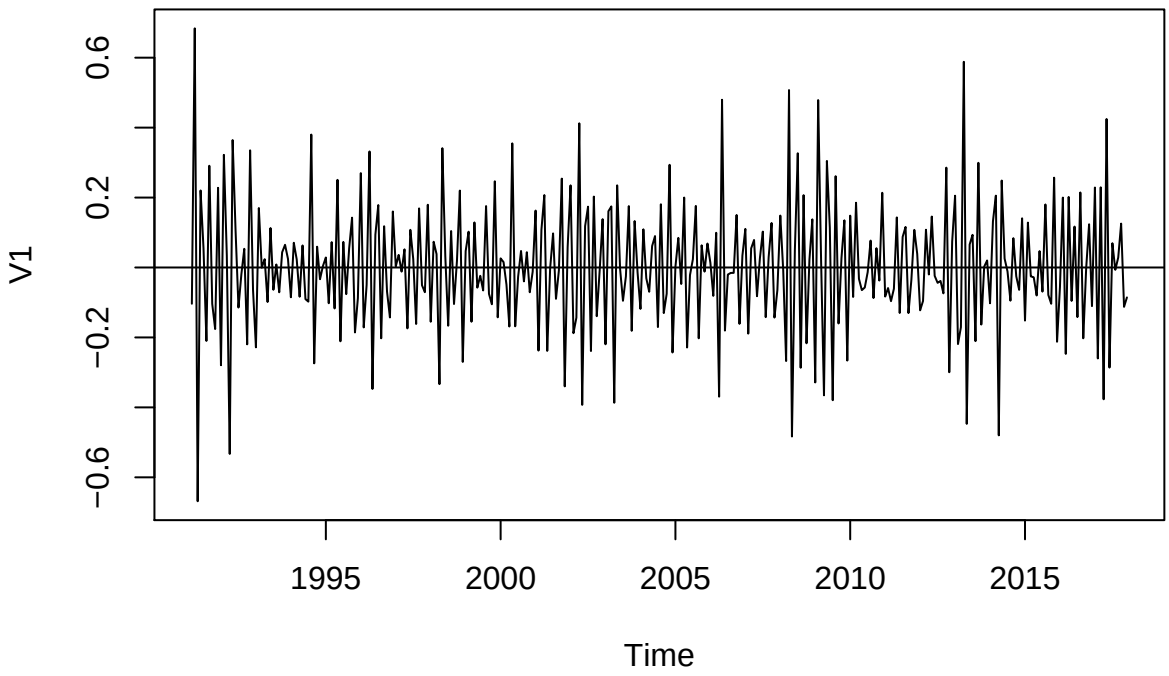


Estudi de la mitjana

Després d'aplicar la diferenciació estacional a la nostra sèrie, veiem que la mitjana no és constant, com podem veure clarament a l'outlier del 2008 i els següents mesos, de manera que aplicarem també una diferenciació regular.



Per tal d'assegurar que realment la mitjana és constant, apliquem una altra diferenciació regular.



Tria de la millor transformació

Comparem les variàncies dels diferents tractaments de la sèrie per veure quina és la millor opció, que serà aquella amb menor variància.

```
##         V1
## V1 1415351
```

```
##           V1
## V1 0.2786358
```

```
##            V1
## V1 0.03053649
```

```
##            V1
## V1 0.01193678
```

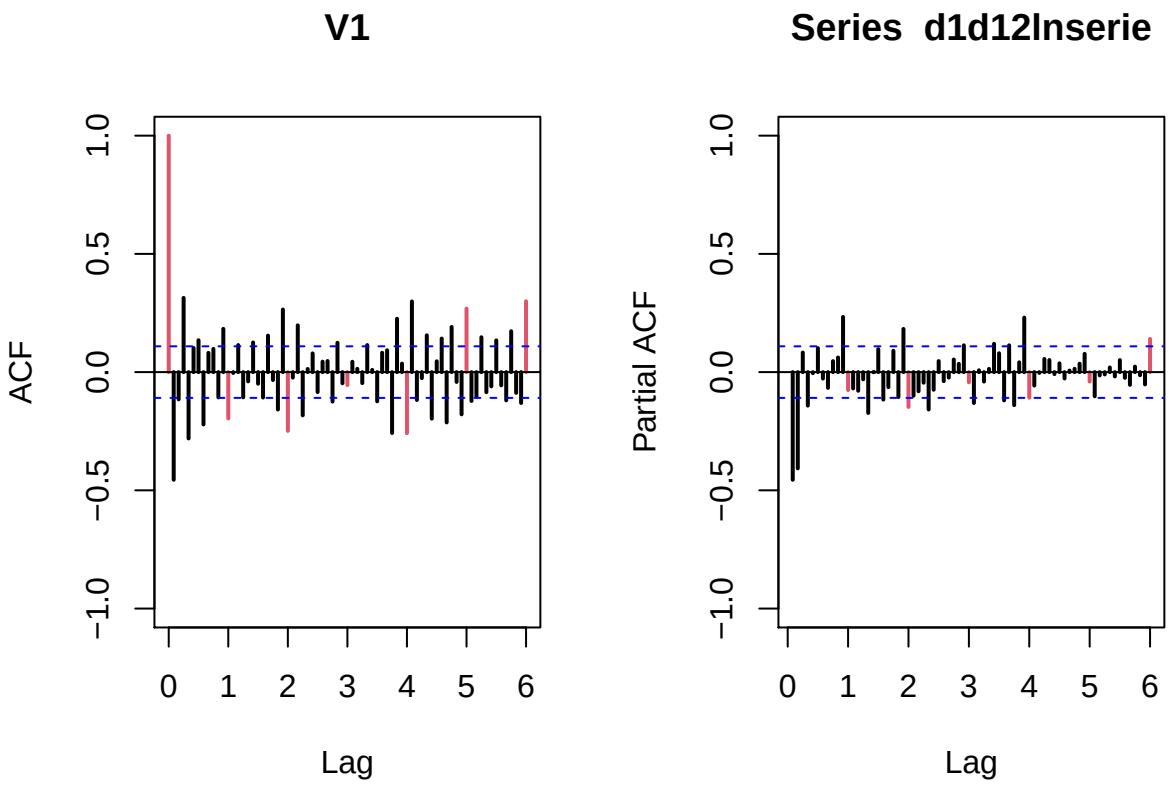
```
##            V1
## V1 0.03479183
```

Veiem que el millor tractament per a la nostra sèrie és una diferenciació estacional i un regular, per tant, treballarem amb la sèrie "d1d12lnserie".

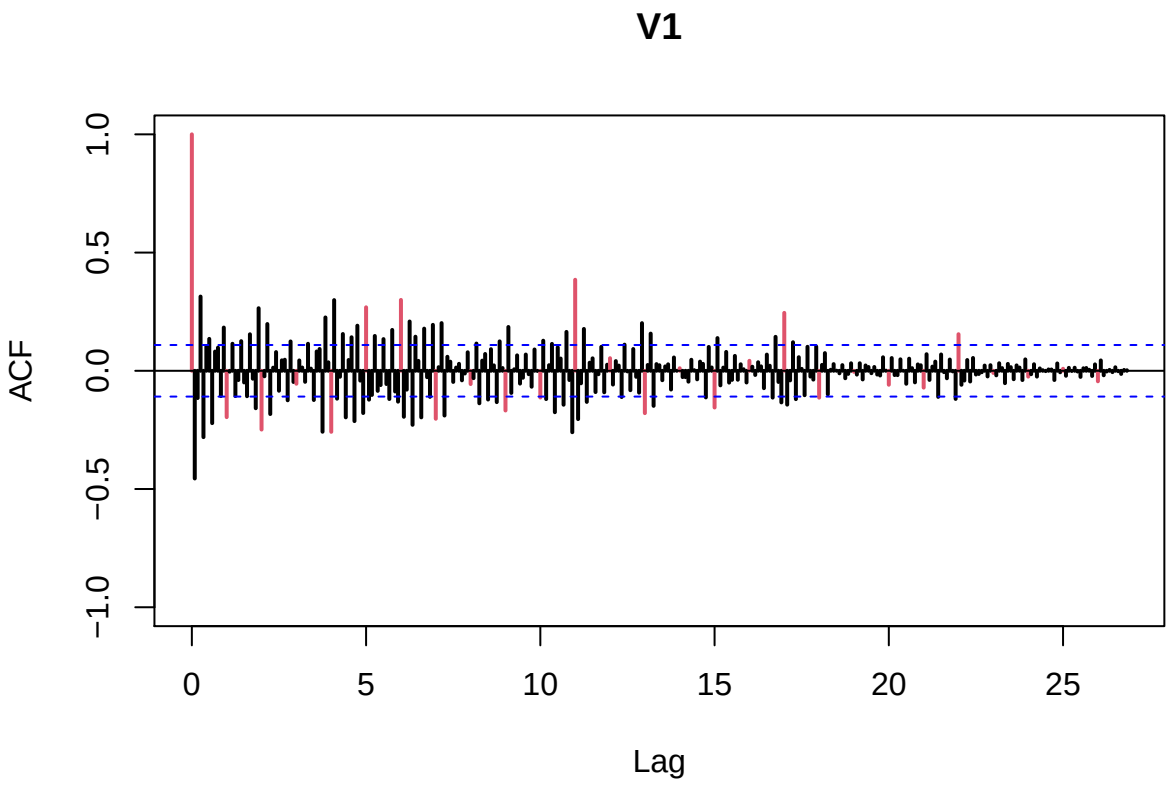
$$(1-B)(1-B^{12})log(X_t) = Z_t$$

Identificació de models

Per poder identifcar els models, representem i analitzarem l'ACF i la PACF de la nostra sèrie transformada:



Veient això, sembla ser que l'ACF no presenta un comportament decreixent en la part que estem observem, per tant, ampliem nombre de retards per veure si realmente decreix al llarg del temps.



Efectivament, l'ACF presenta un patró de decreixement. Amb això i el que havíem vist al PACF podríem plantejar els següents models:

- $Model\ 1 = ARIMA(4,1,0)(0,1,2)_{12}$
- $Model\ 2 = ARIMA(1,1,1)(1,1,1)_{12}$

En el primer model, com veiem que hi ha dos pics molts clars al PACF identifiquem un model AR a la part regular, a més del decreixement de l'ACF. Així, q = 0 i com veiem que el quart retard sembla significatiu, decidim agafar p = 4, ja que sempre és millor sobreparametritzar d'entrada i anar treient paràmetres durant l'estimació de models. Per la part estacional, veiem dos retards significatius a l'ACF, i el tercer ja no ho és, de manera que triem un model MA, on P = 0, amb Q = 2. Tot i això, després sembla haver més retards significatius, però amb l'estimació del model comprovarem si realment aquest model s'ajusta al comportament de la sèrie. D'altra banda, la sèrie presenta una diferenciació estacional i una regular, per tant, d = D = 1.

Pel segon model, utilitzarem un model ARMA(1,1) tant per a la part regular com per a l'estacional, ja que podem observar patrons de decreixement prou clars als dos gràfics. Per tant, tot i que el model ARMA(1,1) presenta pitjors propietats, pot explicar bé la sèrie. Així que p = q = P = Q = 1, i com la sèrie presenta una diferenciació estacional i una regular, d = D = 1.

Estimació dels models

Model 1

Comencem fent un test ràtio dels coeficients del primer model per comprovar que són estadísticament significatius (prenem d = D = 0 perquè ja hem fet les diferenciacions de la sèrie prèviament).

```
##
## Call:
## arima(x = d1d12lnserie, order = c(4, 0, 0), seasonal = list(order = c(0, 0,
##     2), period = 12))
##
## Coefficients:
##           ar1      ar2     ar3      ar4     sma1     sma2  intercept
##       -0.5374  -0.3381  0.0806  -0.1160  -0.6307  -0.3693     -3e-04
## s.e.   0.0555   0.0633  0.0646   0.0566   0.0751   0.0643      3e-04
##
## sigma^2 estimated as 0.004683:  log likelihood = 390.39,  aic = -764.78
```

Podem veure que, estadísticament, el terme independent val 0, i el seu test ràtio és menor que 2 en valor absolut, per tant, al nostre model no és important i no cal tenir-la en compte.

```
##
## Call:
## arima(x = lnserie, order = c(4, 1, 0), seasonal = list(order = c(0, 1, 2), period = 12))
##
## Coefficients:
##           ar1      ar2     ar3      ar4     sma1     sma2
##       -0.5353  -0.3339  0.0854  -0.1129  -0.6276  -0.3724
## s.e.   0.0556   0.0632  0.0646   0.0566   0.0773   0.0645
##
## sigma^2 estimated as 0.004696:  log likelihood = 389.96,  aic = -765.91
```

Veiem que l'AIC ha disminuit, de manera que aquest model explica millor la nostra sèrie i a més utilitza menys paràmetres. Fent el ràtio test dels altres paràmetres, veiem que el tercer de la part AR de la diferenciació regular té un valor menor que 2, de manera que provem a eliminar-lo.

```
## Warning in arima(lnserie, order = c(4, 1, 0), seasonal = list(order = c(0, :
## some AR parameters were fixed: setting transform.pars = FALSE
```

```
##
## Call:
## arima(x = lnserie, order = c(4, 1, 0), seasonal = list(order = c(0, 1, 2), period = 12),
##     fixed = c(NA, NA, 0, NA, NA, NA))
##
## Coefficients:
##           ar1      ar2  ar3      ar4     sma1     sma2
##       -0.5559  -0.3839    0  -0.1497  -0.6117  -0.3883
## s.e.   0.0534   0.0508    0   0.0494   0.0770   0.0630
##
## sigma^2 estimated as 0.004723:  log likelihood = 389.08,  aic = -766.17
```

Veiem que efectivament l'AIC disminueix, i com que els altres test ràtio són bastant superiors a 2, ens quedarem amb aquesta estimació del primer model.

```
## Warning in arima(lnserie, order = c(4, 1, 0), seasonal = list(order = c(0, :
## some AR parameters were fixed: setting transform.pars = FALSE
```

```
##
## Call:
## arima(x = lnserie, order = c(4, 1, 0), seasonal = list(order = c(0, 1, 2), period = 12),
##     fixed = c(NA, NA, 0, NA, NA, NA))
##
## Coefficients:
##           ar1      ar2  ar3      ar4     sma1     sma2
##       -0.5559  -0.3839    0  -0.1497  -0.6117  -0.3883
## s.e.   0.0534   0.0508    0   0.0494   0.0770   0.0630
##
## sigma^2 estimated as 0.004723:  log likelihood = 389.08,  aic = -766.17
```

Model 2

Procedirem ara d'igual manera amb el segon model.

```
##
## Call:
## arima(x = d1d12lnserie, order = c(1, 0, 1), seasonal = list(order = c(1, 0,
##     1), period = 12))
##
## Coefficients:
##           ar1      ma1    sar1     sma1  intercept
##       -0.1426  -0.4562  0.2967  -1.0000     -3e-04
## s.e.   0.0832   0.0684  0.0575   0.0374      3e-04
##
## sigma^2 estimated as 0.0054:  log likelihood = 368.27,  aic = -724.53
```

Podem veure que, estadísticament, el terme independent val 0, i el seu test ràtio és menor que 2 en valor absolut, per tant, al nostre model no és important i no cal tenir-la en compte.

```
##
## Call:
## arima(x = lnserie, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 1), period = 12))
##
## Coefficients:
##           ar1      ma1    sar1     sma1
##       -0.1453  -0.4517  0.3026  -1.0000
## s.e.   0.0830   0.0683  0.0574   0.0405
##
## sigma^2 estimated as 0.005418:  log likelihood = 367.8,  aic = -725.59
```

Veiem que l'AIC ha disminuit, de manera que aquest model explica millor la nostra sèrie i a més utilitza menys paràmetres. Ara provarem a augmentar els paràmetres d'AR i MA de les parts regular i estacional per veure si millora el model.

```
## [1] -758.797
```

Veiem que augmentant afegint un paràmetre a l'AR de la part regular l'AIC disminueix, per tant, el model millora. Per tant, provem si afegir un altre manté aquest efecte.

```
## [1] -756.8886
```

Veiem que ara empitjora el model, per tant, provarem a afegir altres paràmetres.

```
## [1] -756.9934
```

Veiem que afegint paràmetres al MA de la part regular el model no millora, per tant, passarem a provar a afegir paràmetres a la part estacional del model.

```
## [1] -766.144
```

Ara sí que disminueix l'AIC i, en conseqüència, el model millora, per tant, provarem a afegir un altre paràmetre a aquest AR de la part estacional del model.

```
## [1] -764.144
```

Veiem que no millora el model, per tant, passem a provar afegir un paràmetre al MA de la part estacional.

```
## [1] -764.5248
```

Veiem que tampoc, millora, per tant, ens quedem amb el model al qual hem arribat afegint un paràmetre a l'AR de cadascuna de les parts del model.

Ara, provarem a treure variables per evitar la sobreparametrització del model, seguint el principi de parismònia.

```
## [1] -704.3111
```

```
## [1] -758.9944
```

```
## [1] -696.6604
```

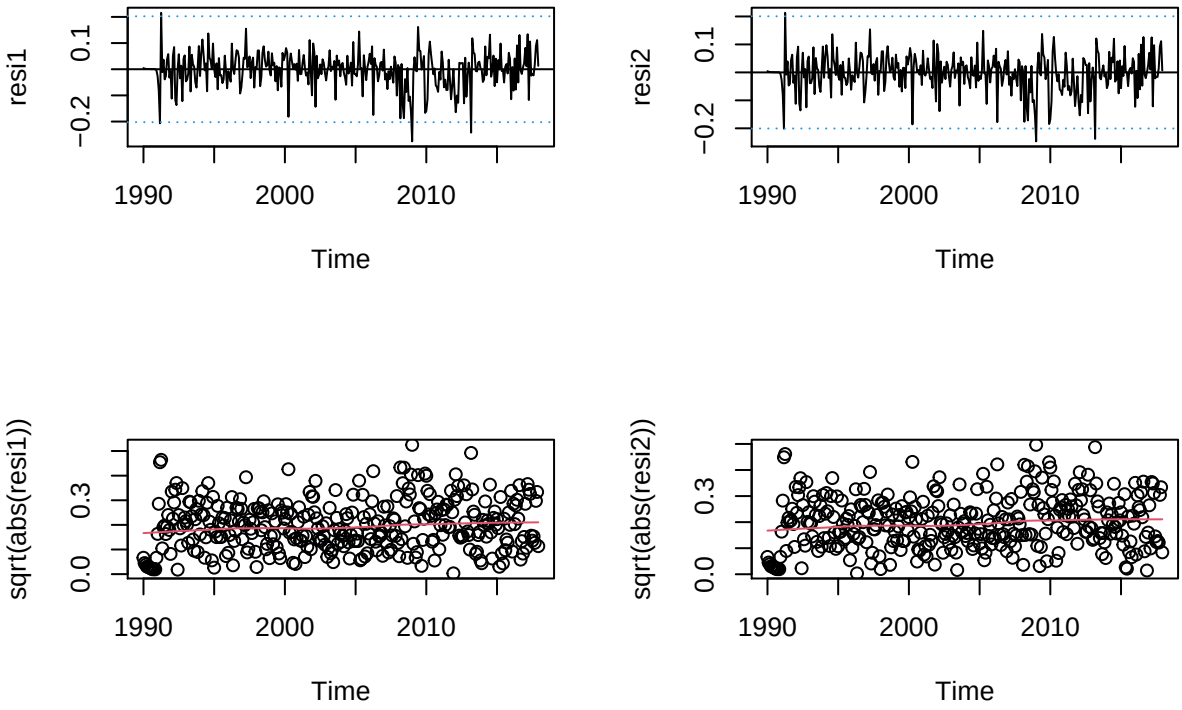
Veiem que en cap cas disminueix l'AIC, per tant, no millora el model que ja teníem, que és amb el que ens quedarem.

```
##
## Call:
## arima(x = lnserie, order = c(2, 1, 1), seasonal = list(order = c(2, 1, 1), period = 12))
##
## Coefficients:
##           ar1      ar2     ma1    sar1     sar2     sma1
##       -0.9360  -0.5196  0.4105  0.3206  -0.1854  -1.0000
## s.e.   0.0961   0.0543  0.1070  0.0573   0.0597   0.0795
##
## sigma^2 estimated as 0.004639:  log likelihood = 390.07,  aic = -766.14
```

Validació del model

Variància constant

Per començar la validació dels nostres models, començarem fent els plots dels residus i de l'arrel quadrada dels valors absoluts dels residus. Amb això, podrem detectar outliers, comprovarem que els models compleixin el principi d'homocedasticitat i veurem si la seva variància és constant.



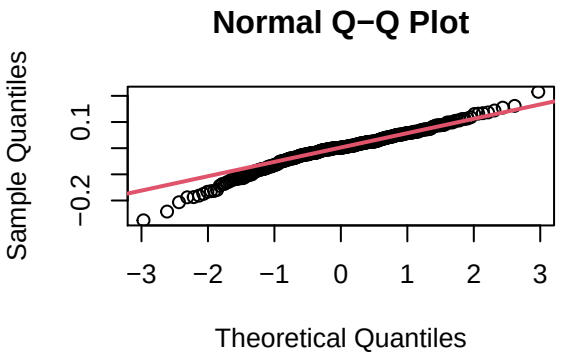
Si mirem els dos primers gràfics dels residus, podem veure a simple vista que la variància sembla constant per a ambdós models, amb algunes zones on hi ha més dubte, com a prop del 2008, degut a l'efecte dels outliers.

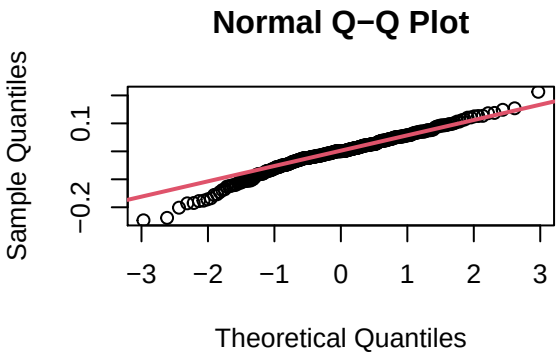
Si mirem ara els altres dos gràfics, veiem que la línea és gairebé recta, símbol de que la variància és constant. Sembla ser, a més a més, que pel model 1 la línia és lleugerament més recta.

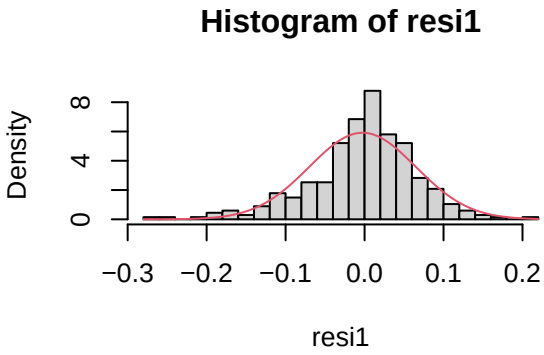
Per tant, en base al que hem vist a aquests gràfics, no podem rebutjar la hipòtesi de la variància constant.

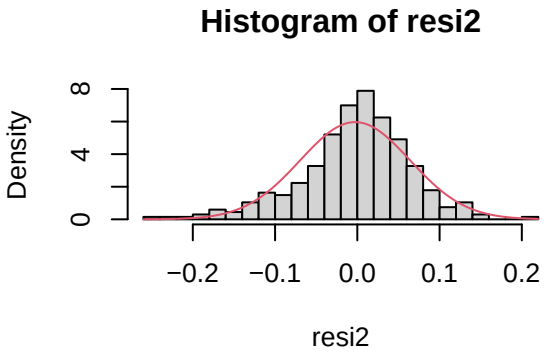
Normalitat

Per veure si els residus provenen d'una distribució normal, farem els plots de normalitat i els histogrames amb la corba normal superposada pels dos models de la sèrie.









Dels dos plots de normalitat, veiem que els residus s'ajusten a la recta a excepció de valors inferiors a -1, on s'allunyen cada cop més, probablement com a conseqüència de la presència d'outliers.

En els histogrames, podem apreciar que el comportament dels residus dels models és similar al d'una distribució normal però desplaçat lleugerament cap a la dreta, com a possible efecte dels outliers, un cop més.

Així, en primera instància, rebutjaríem la hipòtesi de normalitat, però podem aplicar el test de Shapiro-Wilks als residus per acabar de confirmar aquests resultats.

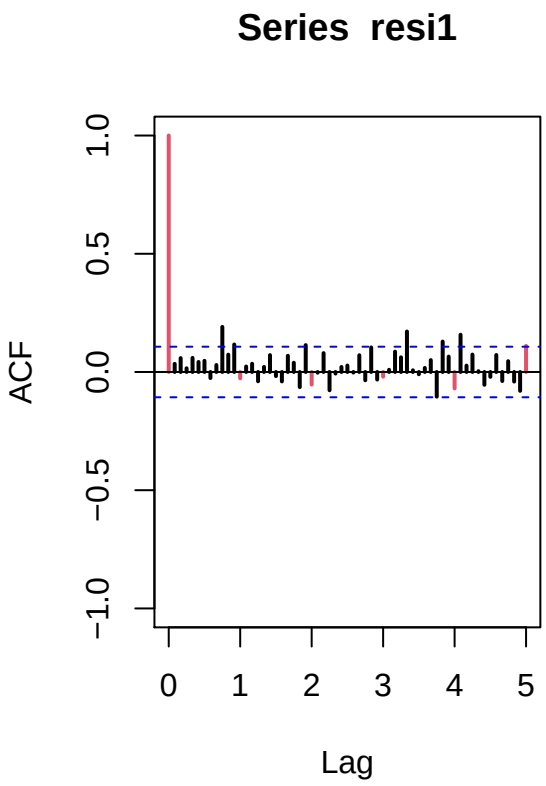
```
##
##  Shapiro-Wilk normality test
##
## data:  resi1
## W = 0.97173, p-value = 3.816e-06
```

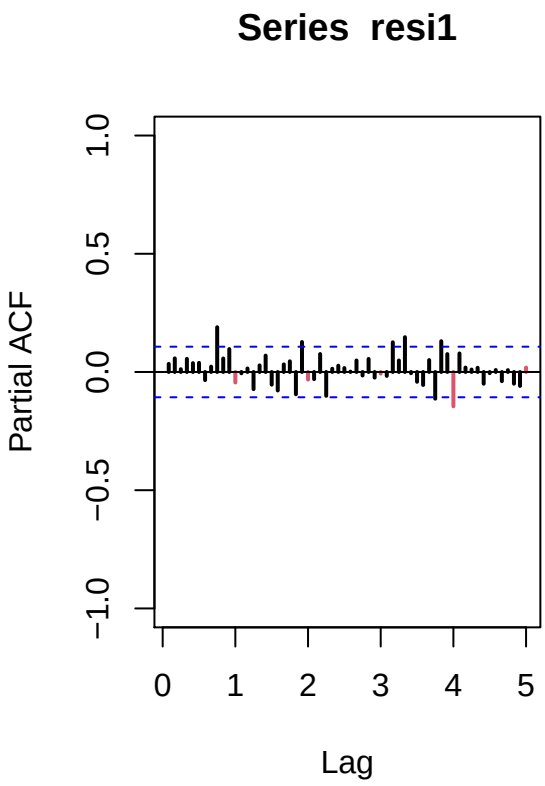
```
##
##  Shapiro-Wilk normality test
##
## data:  resi2
## W = 0.97686, p-value = 3.128e-05
```

El test de Shapiro-Wilks presenta una gran sensibilitat davant la presència de valors atípics, però veient que els p-valors són molt més petits que 0.05 per a ambdós casos, i després d'haver vist els gràfics anteriors, podríem donar el test per bo i, d'aquesta manera, rebutjar la hipòtesi de normalitat de forma definitiva. La normalitat potser la podríem aconseguir tractant els outliers, que és el que farem més endavant.

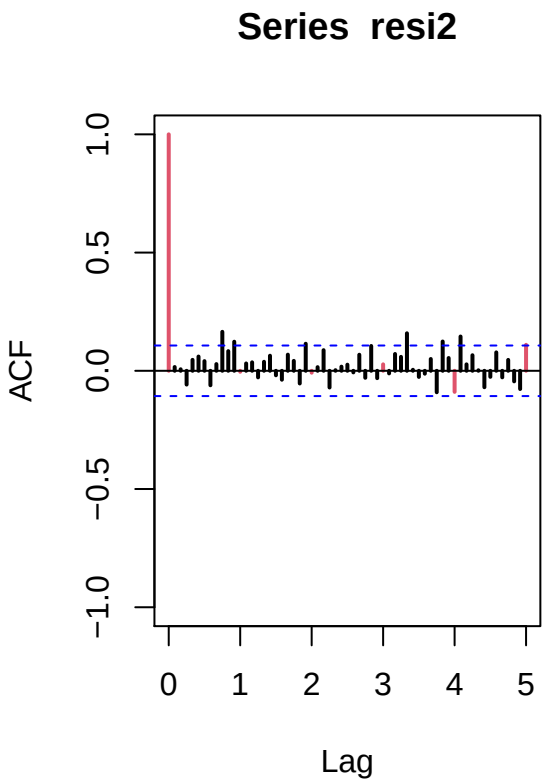
Independència dels residus

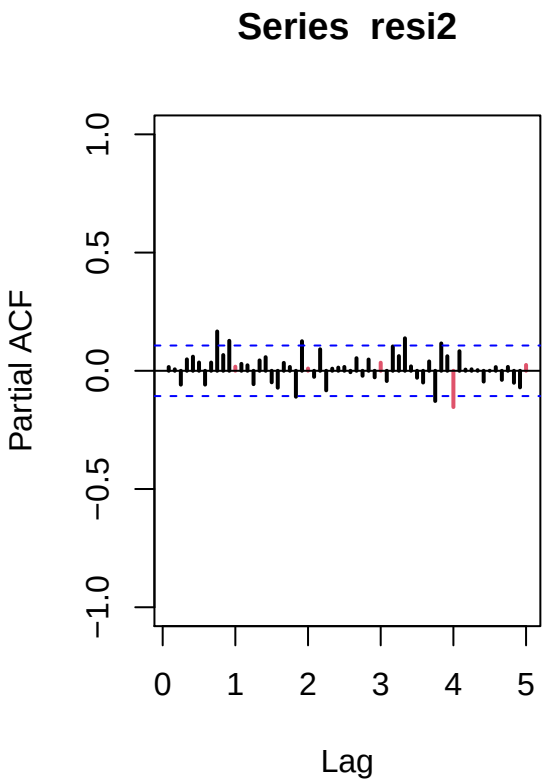
Farem ara els gràfics de l'ACF i el PACF dels residus dels dos models per comprovar la independència d'aquests.





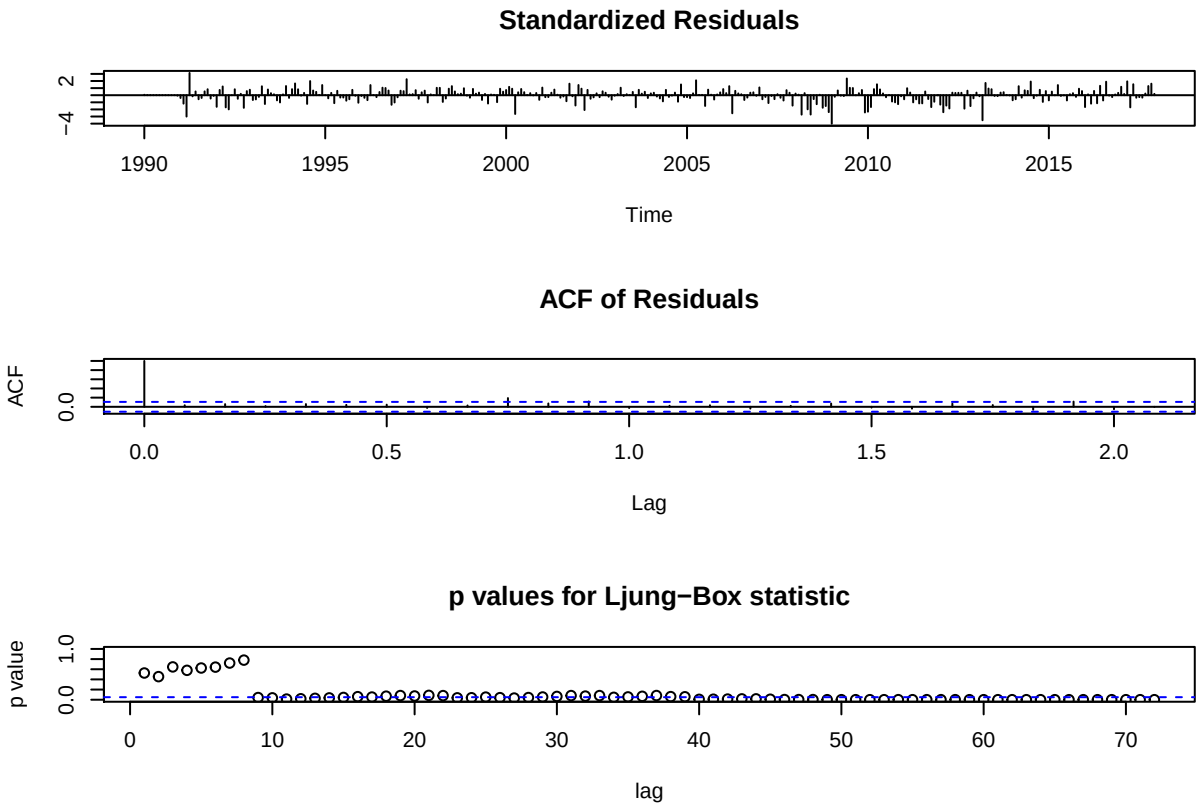
Pel primer model, podem veure alguns valors al llarg de la sèrie fora de l'interval de confiança de l'ACF i el PACF, fet que ens fa sospitar que els residus no són independents. Així, no podem identificar l'ACF i el PACF amb soroll blanc, deixant una part de la variabilitat de les dades del nostre model sense explicar.



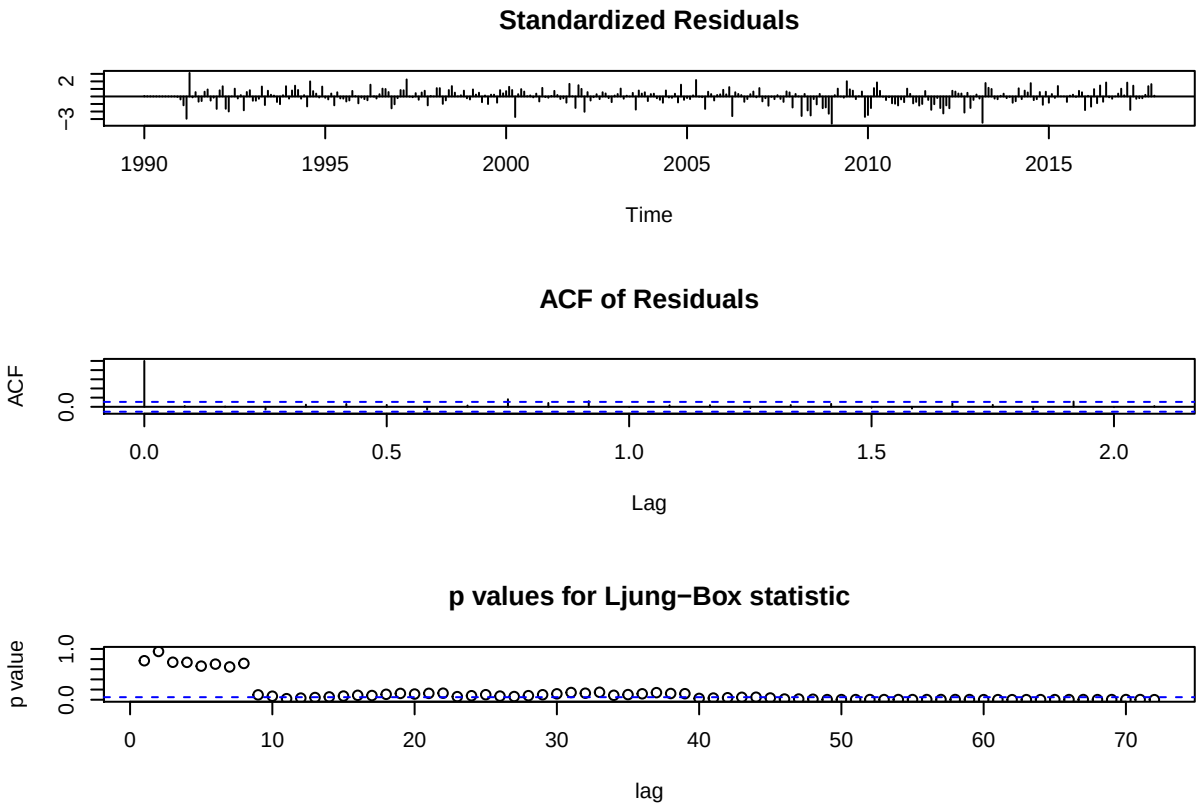


Per aquest segon model, podem veure alguns valors al llarg de la sèrie fora de l'interval de confiança de l'ACF i el PACF, fet que ens fa sospitar que els residus no són independents. Així, no podem identificar l'ACF i el PACF amb soroll blanc, deixant una part de la variabilitat de les dades del nostre model sense explicar.

Anem a veure ara la representació dels p-valors pel test de Ljung-Box per tal de confirmar el que hem vist:



Veiem pel primer model que, a partir dels retards que havíem vist a l'ACF i al PACF com a significatius, trobem valors del p-valor per sota del 0.05, rebutjant així la hipòtesi nul · la d'independència entre els residus.



Veiem el mateix comportament per aquest segon model, on a partir dels retards que havíem vist a l'ACF i al PACF com a significatius, trobem valors del p-valor per sota del 0.05, rebutjant així la hipòtesi nul · la d'independència entre els residus.

Aquesta dependència entre els residus pot ser conseqüència de la volatibilitat de les dades, que ens porta a pensar que hi ha una part d'aquestes que no s'expliquen als nostres models, que es basen en el passat.

Estacionarietat i invertabilitat

Per expressar els models com a AR i MA infinits, necessitem saber les arrels dels polinomis característics de ϕ i θ, respectivament. A més, analitzant-los, podrem dir si es tracta de models invertibles i estacionaris.

Perquè es compleixi la estabilitat s'ha de conseguir que sigui causal, és a dir, que les arrels del polinomi característic AR (coeficients ϕ) siguin més grans que 1; i invertible, es a dir, que les arrels del polinomi característic MA (coeficients θ) siguin més grans que 1.

```
## [1] 1.965614 1.315044 1.315044 1.965614
```

```
##  [1] 1.082033 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
##  [9] 1.000000 1.082033 1.082033 1.000000 1.000000 1.000000 1.000000 1.082033
## [17] 1.082033 1.082033 1.082033 1.082033 1.082033 1.082033 1.082033 1.082033
```

Com podem veure, tant la part AR com la MA d'aquest primer model presenta totes les seves arrels més grans que 1, per tant, podem assegurar que és causal (arrels AR>1.2) i invertible (arrels MA>1), de manera que tindrem bones propietats als intervals de confiança i estabilitat al model.

```
##  [1] 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736
##  [9] 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736
## [17] 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736 1.072736
## [25] 1.387288 1.387288
```

```
##  [1] 1.000001 1.000001 1.000001 1.000001 1.000001 1.000001 1.000001 1.000001
##  [9] 1.000001 1.000001 1.000001 1.000001 2.435828
```

Com podem veure, tant la part AR com la MA d'aquest segon model presenta totes les seves arrels més grans que 1, però no gaire. Pel cas de la invertibilitat sí que podríem considerar que aquest model ho és (arrels MA>1), però en quant a l'estabilitat, les arrels de la part AR són bastant properes a 1(< 1.2). Com per modelitzar necessitem que el model sigui causal (ho imposa el paquet que utilitzem), de manera artificial sempre generarà un valor per sobre de 1, encara que sigui molt proper. El llindar per considerar si es causal o no és amb una arrel per sobre o per sota de 1.2, que en aquest cas no és així. Per tant, no considerem el model causal i no tindrem bones propietats als intervals de confiança ni estabilitat al model.

Així, veiem que el nostre primer model presenta millors propietats que el segon, fet que té sentit, ja que aquest últim és un model ARMA, que es caracteritza per tenir pitjors propietats. En concret, com el primer model és invertible i causal, podem dir que és estable, en canvi, el segon model no ho és.

Mesures d'adequació a les dades

Calculem ara les mesures d'adequació a les dades (AIC i BIC):

```
## [1] -766.1697
```

```
## [1] -766.144
```

```
## [1] -743.5038
```

```
## [1] -739.7004
```

Podem veure que tant el valor de l'AIC com el del BIC són millors pel primer model. En aquest cas, té sentit que el BIC sigui millor pel primer model, ja que utilitza menys paràmetres.

Capacitat de previsió

Per tal de valorar la capacitat de previsió dels models, eliminarem el darrer any de la sèrie original per veure si la previsió s'ajusta als valors originals de la sèrie. Per tant, començarem eliminant les 12 darreres observacions de la sèrie logarítmica.

Ajustarem dos nous models a partir dels que teníem però ara per a la sèrie sense les 12 darreres observacions.

```
## Warning in arima(lnserie2, order = c(4, 1, 0), seasonal = list(order = c(0, :
## some AR parameters were fixed: setting transform.pars = FALSE
```

```
##
## Call:
## arima(x = lnserie2, order = c(4, 1, 0), seasonal = list(order = c(0, 1, 2),
##     period = 12), fixed = c(NA, NA, 0, NA, NA, NA))
##
```

```
## Coefficients:
##           ar1      ar2  ar3      ar4     sma1     sma2
##       -0.5485  -0.3978    0  -0.1599  -0.5889  -0.4111
## s.e.   0.0542   0.0515    0   0.0506   0.1437   0.0794
##
## sigma^2 estimated as 0.004671:  log likelihood = 375.97,  aic = -739.93
```

En el cas d'aquest model, veiem que tots els coeficients són significatius, ja que els seus test ràtio són majors que 2.

```
##
## Call:
## arima(x = lnserie2, order = c(2, 1, 1), seasonal = list(order = c(2, 1, 1),
##     period = 12))
##
## Coefficients:
##           ar1      ar2     ma1    sar1     sar2     sma1
##       -0.9338  -0.5318  0.4121  0.3283  -0.2062  -0.9257
## s.e.   0.0942   0.0545  0.1052  0.0663   0.0670   0.0729
##
## sigma^2 estimated as 0.004832:  log likelihood = 376.67,  aic = -739.35
```

Veiem que per a aquest model també tenim tots els coeficients significatius, de manera que treballarem amb aquests dos.

Ara, procedirem com abans, i comprovarem que aquests models són estables:

```
## [1] 1.924091 1.299665 1.299665 1.924091
```

```
##  [1] 1.076893 1.000001 1.000001 1.000001 1.000001 1.000001 1.000001 1.000001
##  [9] 1.000001 1.076893 1.076893 1.000001 1.000001 1.000001 1.000001 1.076893
## [17] 1.076893 1.076893 1.076893 1.076893 1.076893 1.076893 1.076893 1.076893
```

```
##  [1] 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990
##  [9] 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990
## [17] 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990 1.067990
## [25] 1.371275 1.371275
```

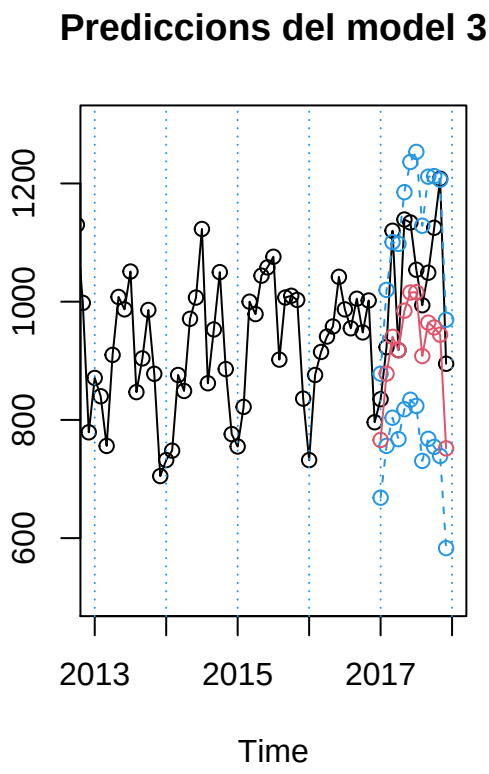
```
##  [1] 1.006451 1.006451 1.006451 1.006451 1.006451 1.006451 1.006451 1.006451
##  [9] 1.006451 1.006451 1.006451 1.006451 2.426406
```

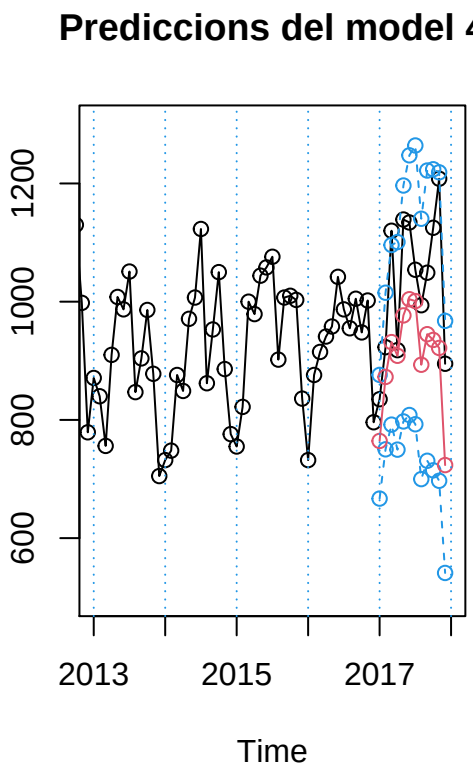
Veiem que, en els dos casos, els models mantenen les mateixes propietats que els de la sèrie original, és a dir, el primer és estable (invertible i causal) però el segon no ("invertible" però no causal).

Així, tenim que els models 1 i 3 són estables i l'2 i el 4 no. Fet que té sentit perquè els models 2 i 4 són ARMA, que són un tipus de models que presenten més inestabilitat.

Amb això comprovat, passarem a utilitzar els models per a la sèrie sense les 12 darreres observacions (models 3 i 4) per fer la predicció i el corresponent interval de confiança al 95% per a l'últim any.

Representem els 5 darrers anys de la sèrie original juntament amb les prediccions i els intervals superposats:





Veiem que les prediccions d'ambdós models són bastant similars però no acaben d'ajustar-se correctament a la sèrie, amb valors de la sèrie original fora de l'interval de confiança. Aquest fet té sentit per la volatibilitat de les dades, però potser tractant els atípics es pot observar una millora.

Calculem ara les mesures de capacitat de previsió (RMSPE i MAPE) a partir de les prediccions puntuals obitngudes prèviament per tal de veure quin dels dos models és el millor per realitzar les previsions.

```
## RMSPE 1: 0.1208058
##  MAPE 1: 0.1052238
##  RMSPE 2: 0.1341049
##  MAPE 2: 0.1185293
```

Podem veure que tots els valors d'aquestes mesures ens han sortit, com intuiem, bastant dolents, fet que indica que els errors són bastant significatius i les nostres prediccions no són gens exactes. En concret, considerem que un error superior al 10% indica un model amb poca capacitat de predicció. Tot i això, veiem que sembla que el model 3 ha fet una millor predicció que el model 4.

Calcularem ara les mitjanes de les amplades dels intervals de confiança de predicció per als dos models.

```
## 371.0933 402.3731
```

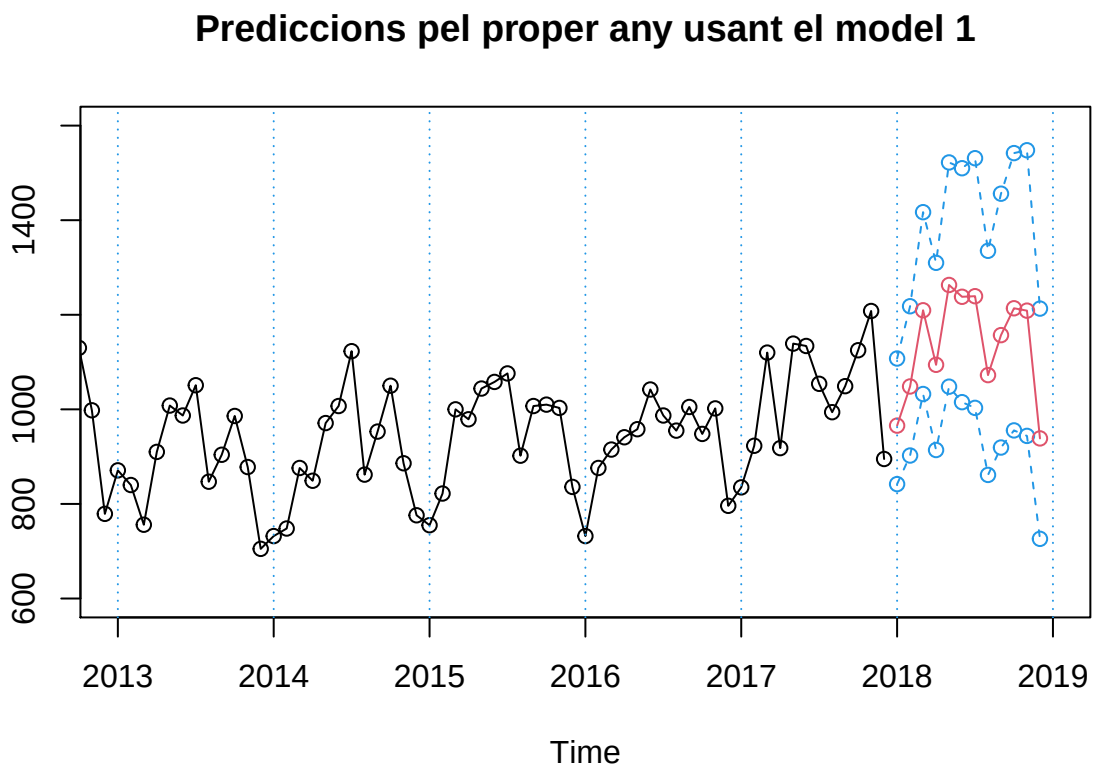
Podem veure que el model 3 té un interval de predicció més petit. A més, com el model 4 és inestable no ens podem fiar del tot d'aquest interval, ja que pot ser més gran del que presenta.

Tria del millor model

Després d'haver treballat aquests dos models proposats hem decidit quedar-nos amb el primer. En primer lloc, hem vist que els dos complien la hipòtesi de variància, però cap dels dos la de normalitat ni la d'independència dels residus, probablement conseqüència de no haver tractat els outliers. A més, tant l'AIC com el BIC eren millors pel primer model. També hem vist que el segon model no era estable, a diferència del primer, que sí es podia considerar com a tal. Tenint en compte això i, tot i la similitud entre els dos models a l'hora de fer les prediccions, la millor capacitat de predicció del primer model, sembla evident triar aquest per a fer les prediccions.

Previsions

Per tant, anem a predir el consum aparent de ciment pel proper any usant el model 1.



Observant els resultats, podem veure que l'interval de confiança es relativament ample en comparacions amb les prediccions anteriors, i en aquest cas contempla l'aparent creixement del consum de ciment al llarg dels següents anys.

Tractament d'atípics

Detecció i interpretació

```
##    Obs type_detected    W_coeff ABS_L_Ratio    Fecha   PercVar
## 1   15            AO -0.2356424    4.663452 Mar 1991  79.00632
## 9   56            LS  0.1356630    3.227689 Ago 1994 114.52959
## 6  184            AO  0.1520371    3.378202 Abr 2005 116.42034
## 11 219            LS -0.1367435    3.357691 Mar 2008  87.21939
## 3  222            LS -0.1954358    4.163614 Jun 2008  82.24761
## 5  227            LS -0.1892381    4.221288 Nov 2008  82.75894
## 4  240            LS -0.1709802    3.716731 Dic 2009  84.28382
## 10 244            LS  0.1326773    3.204527 Abr 2010 114.18814
## 7  266            LS -0.1458176    3.362668 Feb 2012  86.43154
## 2  279            AO -0.2159400    4.394456 Mar 2013  80.57836
## 8  315            AO -0.1453965    3.299206 Mar 2016  86.46794
```

Primerament, podem veure que hi ha un AO (additive outlier) que només afecta a la observació corresponent al març de l'any de 1991. És un factor important en tenir un ABS_L_Ratio gran en comparació als altres outliers i un descens puntual del 21%. No em vist un factor important que afectés a aquesta dada, possiblement hem pensat que sigui que la setmana santa caigués al març del 24-31 i que per les vacances el consum s'aturés durant una setmana (quan acostuma a succeïr a l'abril).

Segonament podem veure un altre AO, amb un increment puntual del 16%. Tampoc veiem una explicació molt clara, però un factor que hem pensat que ha pogut influir és la celebració dels XV Jocs Mediterranis a Almeria, que potser van requerir d'una demanda d'infraestructures més alta del normal i també que la setmana santa caigués al març. Cosa que fés que a l'abril es treballés més.

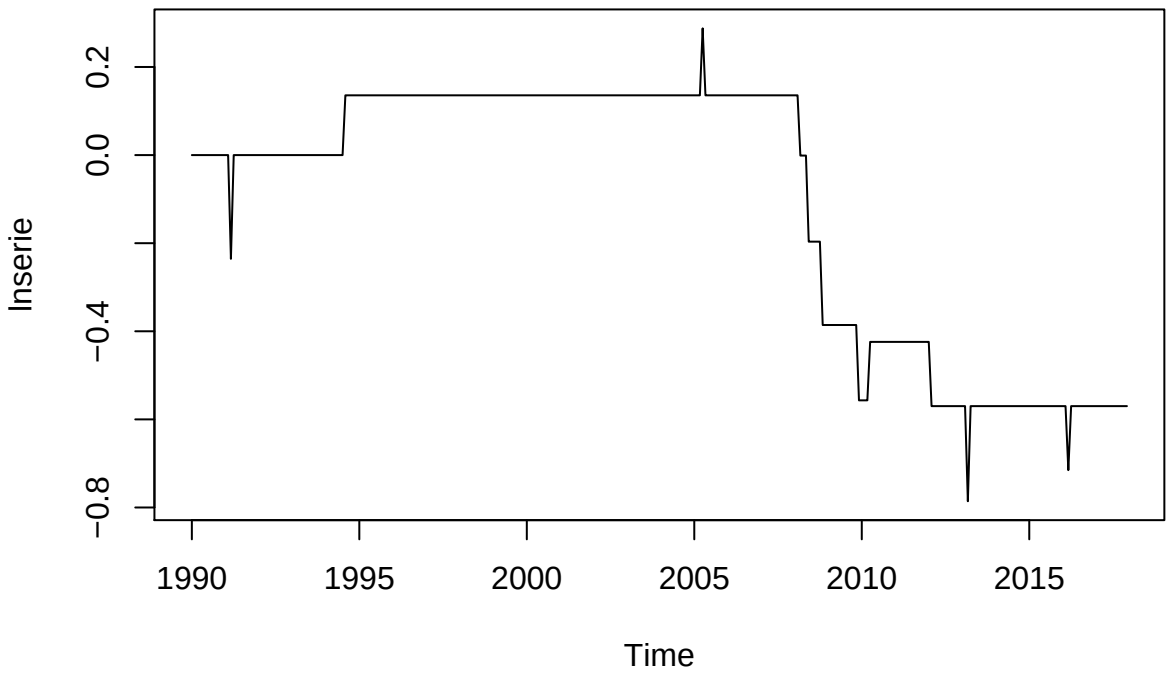
Tercerament, veiem dos LS (levelshift), amb un descens del 18 % en ambdós casos d'una manera continuada fins al dia d'avui (important amb un ABS_L_Ratio gran en comparació als altres outliers). Això va ser resultat de la crisi del 2008 i de la explosió de la "bombolla inmobiliaria" a aquest any, ja que abans hi havia molta expectació amb la construcció i la tinença de propietats. Com els edificis es van devaluar de cop, segurament, es deixaria de contruïr com es feia abans (disminuïnt l'ús de ciment). L'efecte de la bombolla s'allarga al 2009 on veiem un nou LS del 16% de descens i també al Febrer del 2012 amb un altre LS del 14% de descens.

Finalment podem veure que hi ha un AO (additive outlier) que només afecta a la observació corresponent al març de l'any de 2013, es un factor important en tenir un ABS_L_Ratio gran en comparació als altres outliers i un descens puntual del 20%. No em vist un factor important que afectés a aquesta dada, possiblement hem pensat que sigui que la setmana santa caigués al març del 24-31 i que per les vacances el consum s'aturés durant una setmana (quan acostuma a succeïr a l'abril). El mateix passa per 2016, on la setmana santa cau en 20-27 de març, i el descens es de 14%.

Linealització

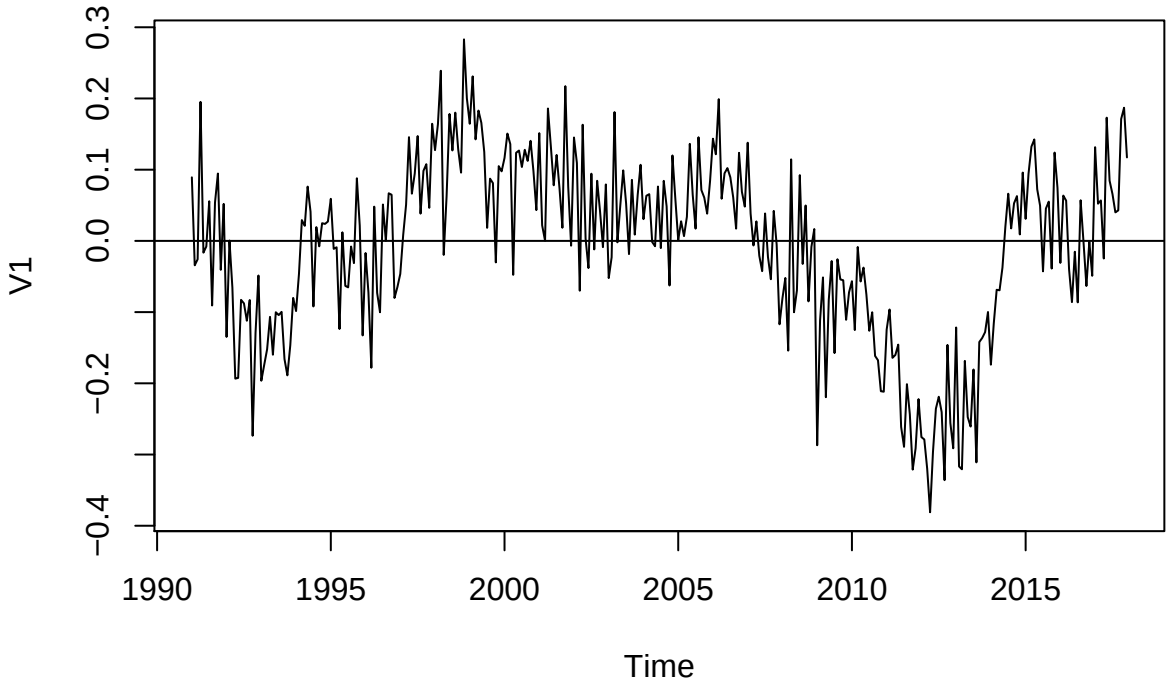
Passem ara a linealitzar la sèrie, és a dir, eliminar l'efecte dels outliers.

Comparem el logaritme de la sèrie original amb el de la sèrie linealitzada.

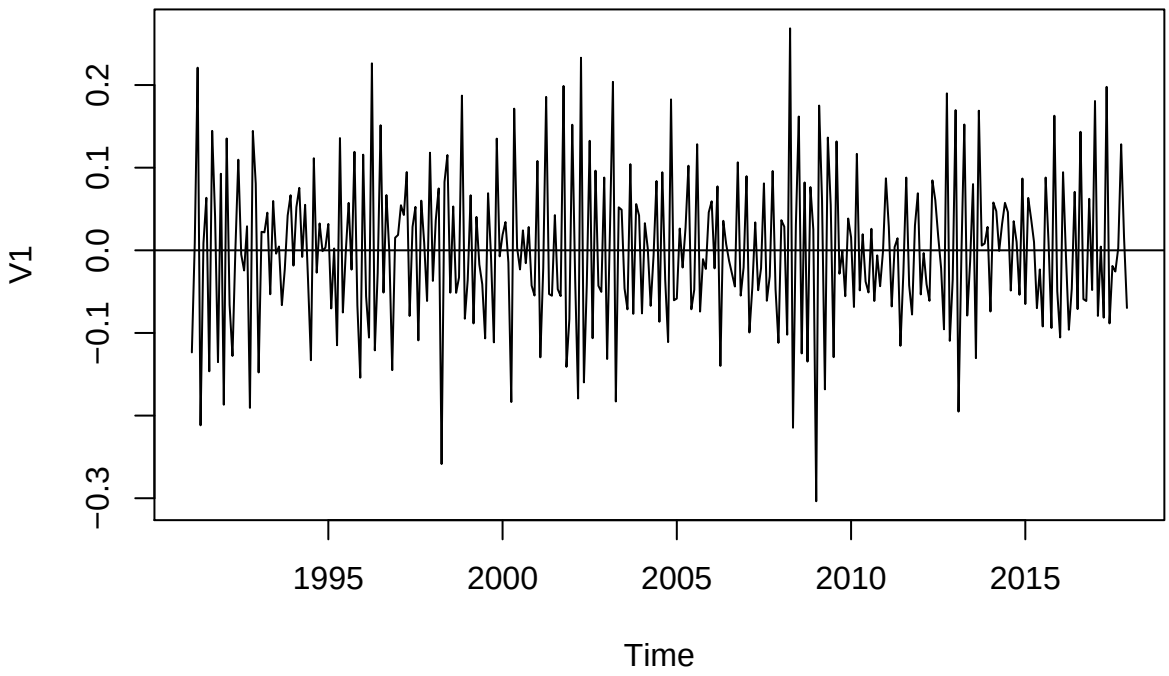


Podem veure clarament l'efecte dels outliers, en aquest cas 4 AO (additive outlier) i 4 LS (level shift), que redueixen els valors del consum de ciment des de l'inici de la sèrie fins al final de la mateixa en un 80%.

Per poder treballar amb la sèrie linealitzada, realitzem les mateixes transformacions que amb la sèrie original, començant per la diferenciació estacional.



Veiem que ara l'efecte dels outliers, en especial el de 2008, s'ha reduït molt, i la sèrie s'apropa molt més a una mitjana constant. Continuem les transformacions amb una diferenciació regular.



Veiem que després d'aquesta transformació la mitjana sembla constant, per tant, ara compararem les variàncies dels diferents tractaments de la sèrie per veure quina és la millor opció, que serà aquella amb menor variància.

```
##        V1
## V1 746474
```

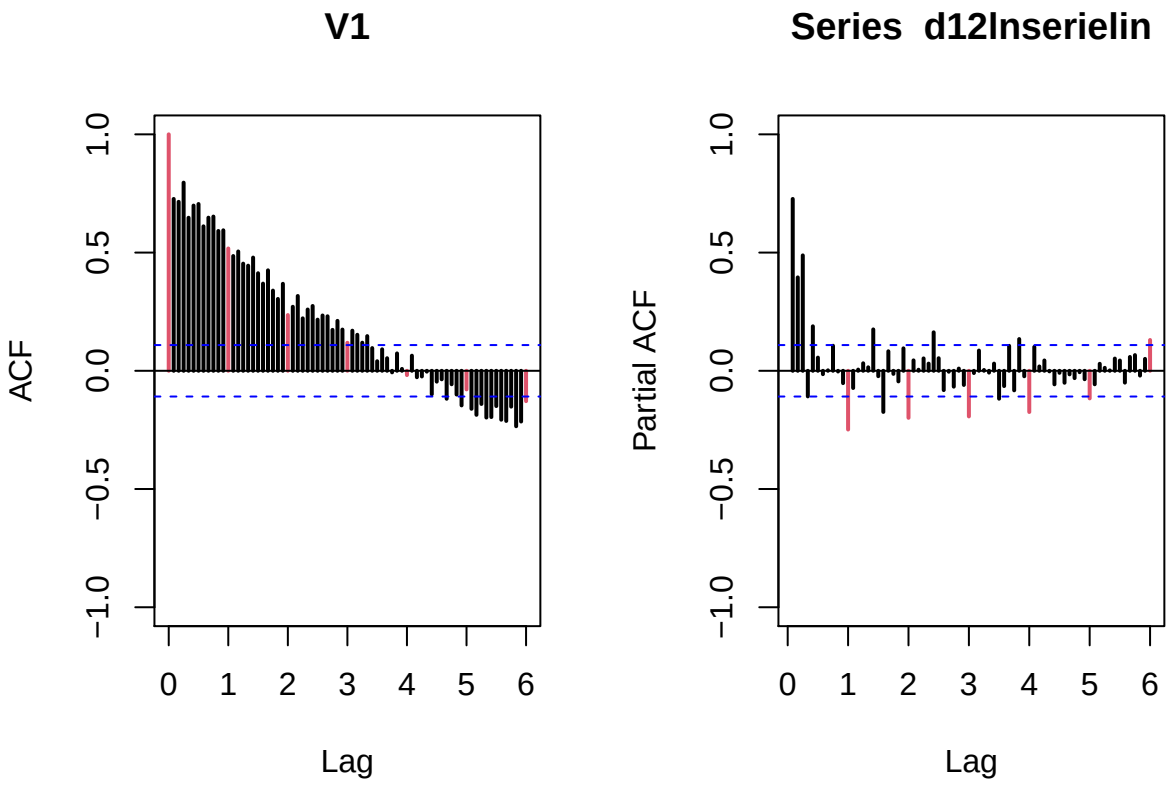
```
##           V1
## V1 0.1082899
```

```
##            V1
## V1 0.01554702
```

```
##             V1
## V1 0.008438042
```

Veiem que el millor tractament per a la nostra sèrie sembla ser una diferenciació estacional i un regular, però com el guany en variància no arriba a ser del doble que només amb la diferenciació estacional, decidim estalviar-nos la diferenciació regular, de manera que prendríem la sèrie amb una única diferenciació estacional (d12lnserielin), aprofitant que ara presenta mitjana constant.

Per poder identifcar els models, representem i analitzarem l'ACF i la PACF de la nostra sèrie transformada:



Analitzant les gràfiques de l'ACF i el PACF, podríem plantejar el següent model AR:

- $Model\ lin = ARIMA(3,0,0)(4,1,0)_{12}$

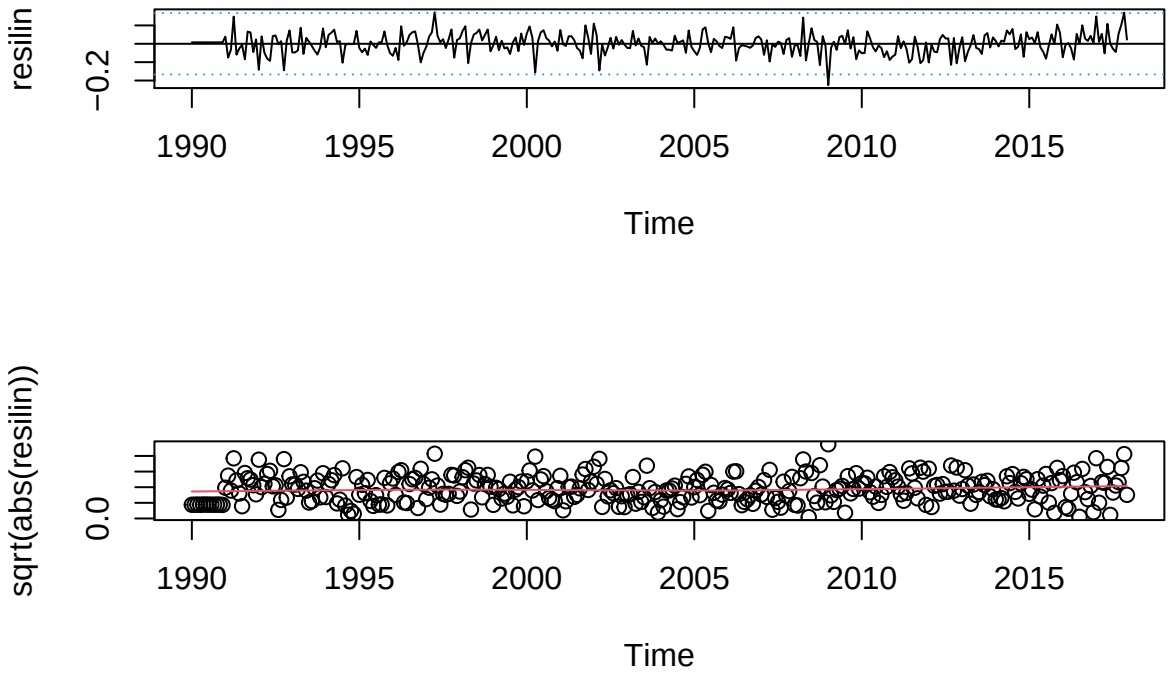
Com veiem 3 pics molt clars al PACF i l'ACF presenta un patró de decreixement, identifiquem un model AR a la part regular, de manera que q = 0 i p = 3. Per la part estacional, veiem 4 retards significatius bastant clars a l'ACF, per tant, Q = 0 i P = 4. D'altra banda, la sèrie presenta només una diferenciació estacional, per tant, d = 0 i D = 1.

Per tal d'estimar el model, ara farem un test ràtio dels coeficients per comprovar que són estadísticament significatius.

```
##
## Call:
## arima(x = lnserielin, order = c(3, 0, 0), seasonal = list(order = c(4, 1, 0),
##     period = 12))
##
## Coefficients:
##          ar1     ar2     ar3     sar1     sar2     sar3     sar4
##       0.3962  0.2149  0.3707  -0.3814  -0.5223  -0.4217  -0.4108
## s.e.  0.0567  0.0554  0.0559   0.0543   0.0569   0.0541   0.0600
##
## sigma^2 estimated as 0.003202:  log likelihood = 462.31,  aic = -908.62
```

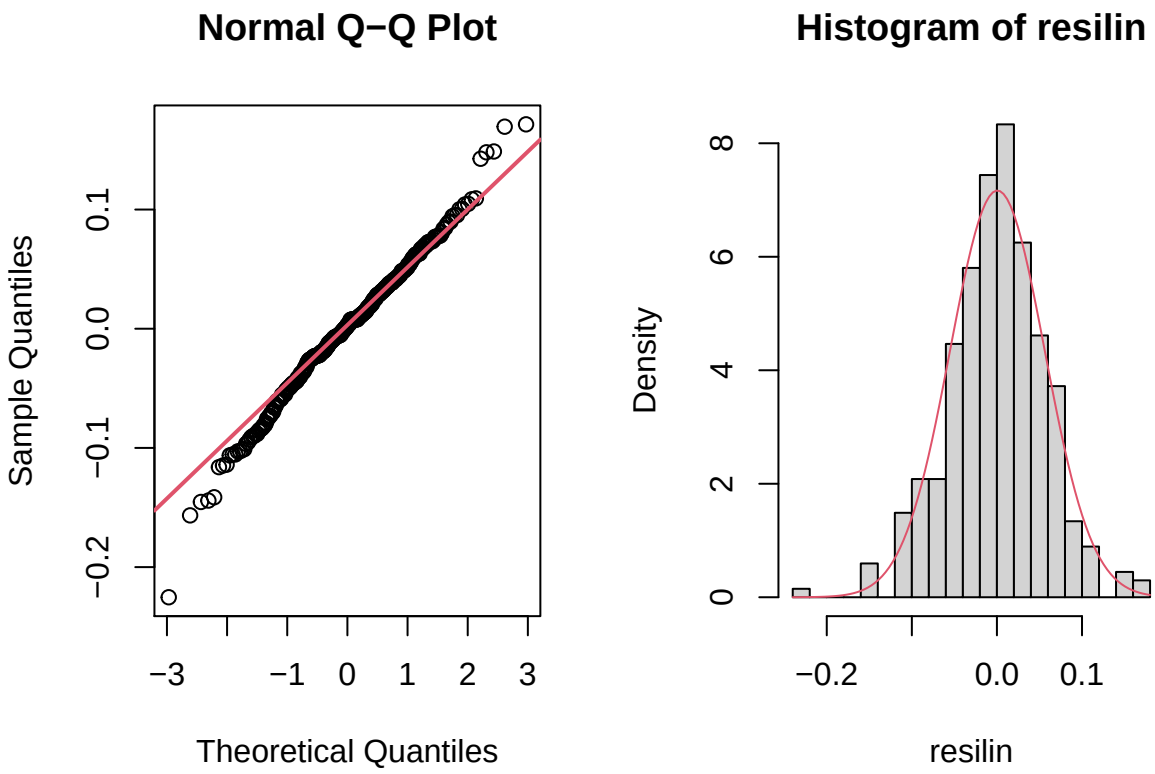
D'entre algunes configuracions que hem provat, aquesta ha estat la que millor explicava el model amb el menor nombre de paràmetres possible, de manera que aquest serà el nostre model. Per tant, passem ara a

validar-lo, començant amb els plots dels residus i de l'arrel quadrada dels valors absoluts dels residus. Amb això comprovarem que els models compleixin el principi d'homocedasticitat i veurem si la seva variància és constant.



Veient els dos gràfics, podem considerar la variància constant, o almenys, no tenim prou evidències com per afirmar el contrari.

Provarem ara que els residus provenen d'una distribució normal, fent el plot de normalitat i l'histograma amb la corba normal superposada.

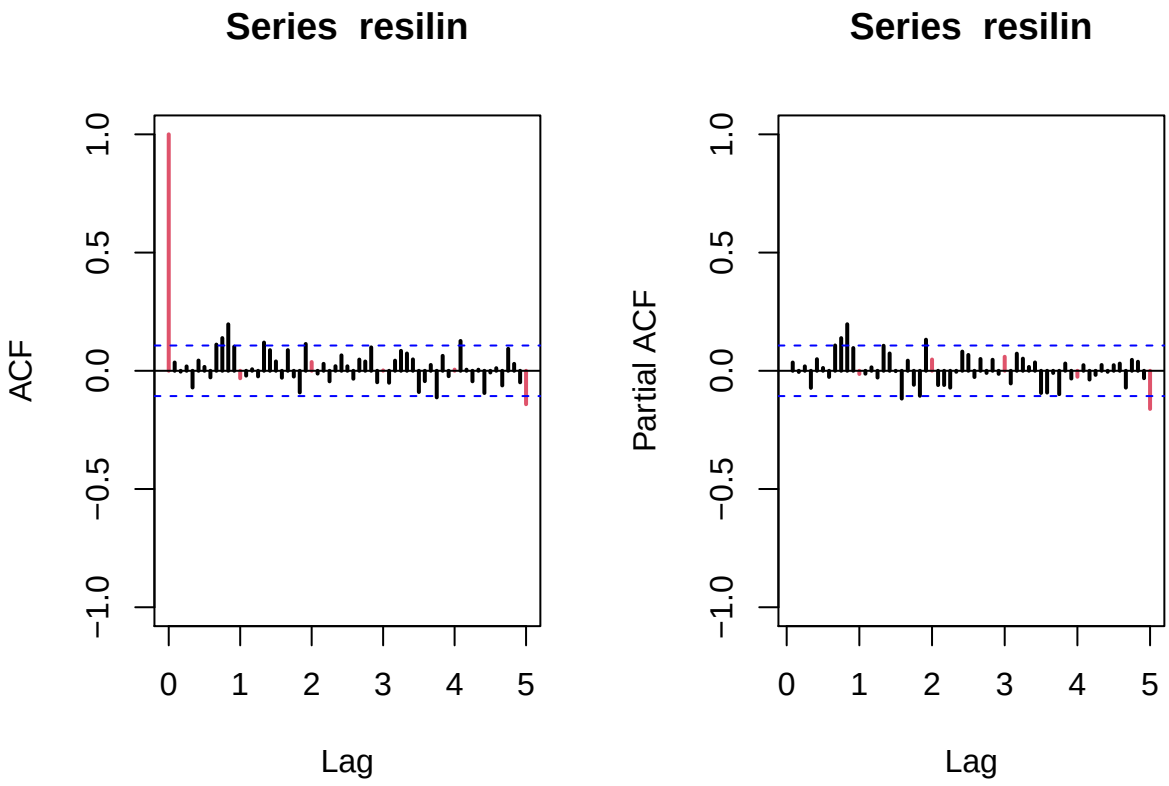


Del plot de normalitat, veiem que els residus s'ajusten en general a la recta, exceptuant algun punt, però no prou evident com per rebutjar l'hipòtesi de normalitat, i a l'histograma veiem un ajust prou adequat, reafirmant les nostres conclusions. De totes formes, podem aplicar el test de Shapiro-Wilks als residus per acabar de confirmar aquests resultats.

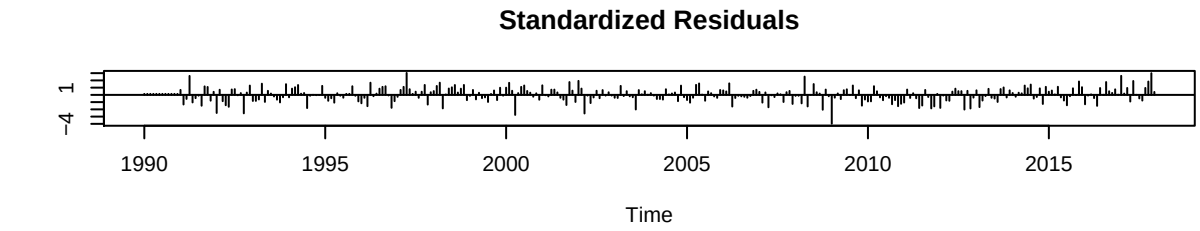
```
##
##  Shapiro-Wilk normality test
##
## data:  resilin
## W = 0.99092, p-value = 0.03638
```

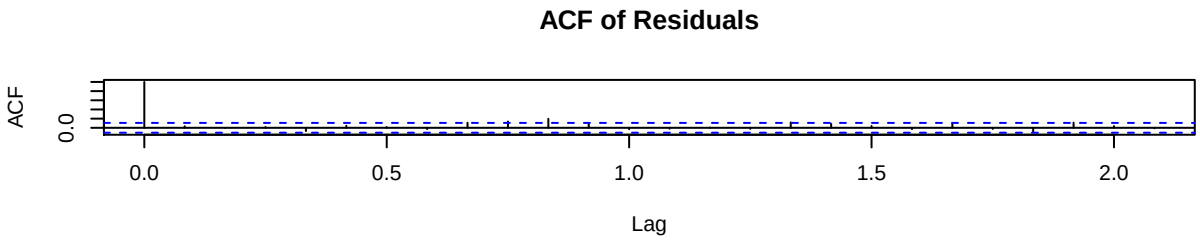
Observem que el test de Shapiro-Wilk té un p-valor que, tot i ser més petit que 0.05, és molt proper. Per tant, tenim en compte l'alta sensibilitat del test, podríem considerar que la hipòtesi de normalitat no es rebutja realment.

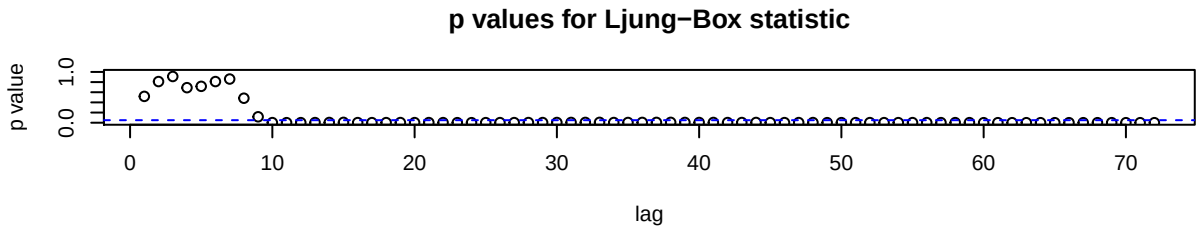
Farem ara els gràfics de l'ACF i el PACF dels residus dels dos models per comprovar la independència d'aquests.



Tot i que aquestes gràfiques s'apropen a una representació de soroll blanc, hi ha algun valor al llarg de la sèrie fora de l'interval de confiança de l'ACF i el PACF, fet que ens fa dubtar de la hipòtesi d'independència dels residus. Anem a veure ara la representació dels p-valors pel test de Ljung-Box per tal de confirmar el que hem vist:







Veiem que, efectivament, trobem valors del p-valor per sota del 0.05, rebutjant així la hipòtesi nul · la d'independència entre els residus. Així, observem que tot i eliminar els outliers hi ha variabilitat dels residus que no es poden explicar mitjançant el temps (el nostre model).

Calculem ara l'AIC del model, tenint en compte que aquesta no té en compte el "pes" d'afegir els paràmetres que representen els outliers, de manera que els afegim nosaltres manualment.

```
## [1] -886.6246
```

Si comparem aquest valor d'AIC amb el que teníem pel model de la sèrie original (model1), veiem que, efectivament, eliminar els outliers permet explicar millor el model.

```
## [1] 120.4549
```

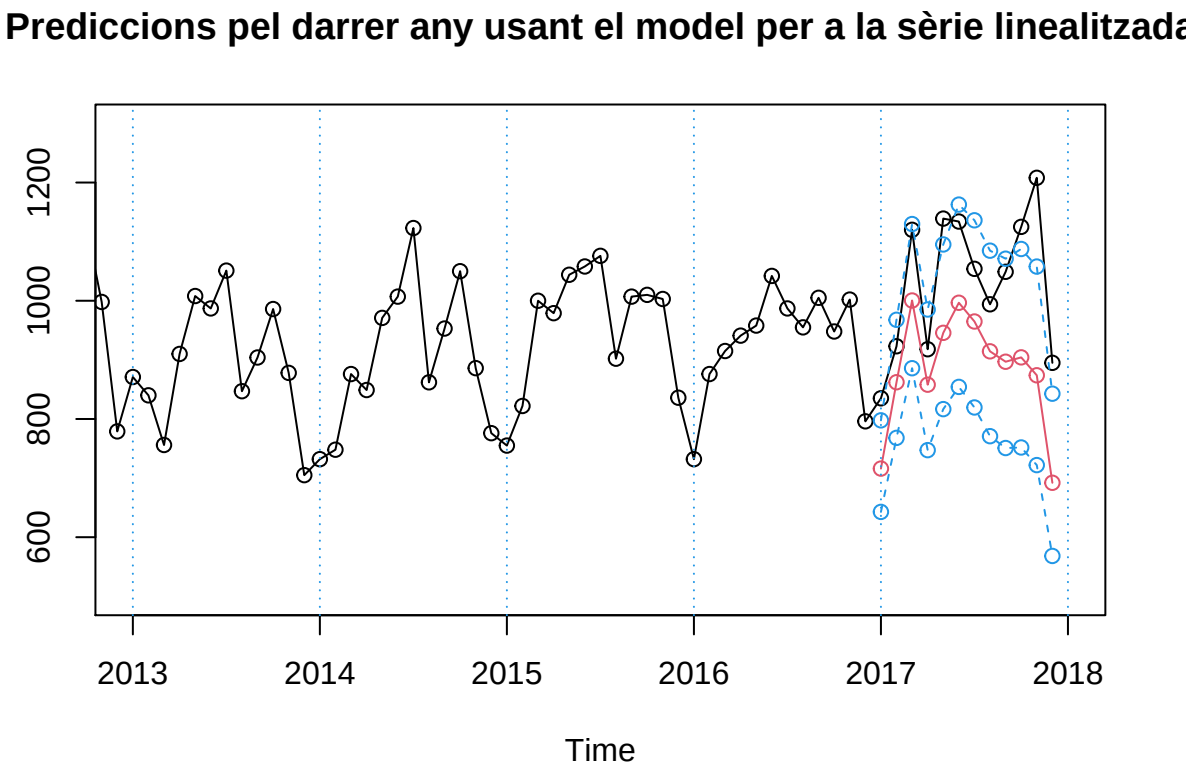
Anem a veure ara la capacitat de predicció d'aquest model utilitzant les observacions de l'últim any com a valors a predir.

Prenem el model amb la sèrie linealitzada sense les observacions dels últims 12 mesos.

```
##
## Call:
## arima(x = lnserielin2, order = c(3, 0, 0), seasonal = list(order = c(4, 1, 0),
##     period = 12))
##
## Coefficients:
##          ar1     ar2     ar3     sar1     sar2     sar3     sar4
```

```
##       0.3837  0.2015  0.3977  -0.3643  -0.5318  -0.4112  -0.4325
## s.e.  0.0566  0.0555  0.0553   0.0539   0.0570   0.0533   0.0594
##
## sigma^2 estimated as 0.003032:  log likelihood = 452.81,  aic = -889.61
```

Passem ara a predir el comportament de la sèrie pels últims 12 mesos utilitzant el model de la sèrie linealitzada.



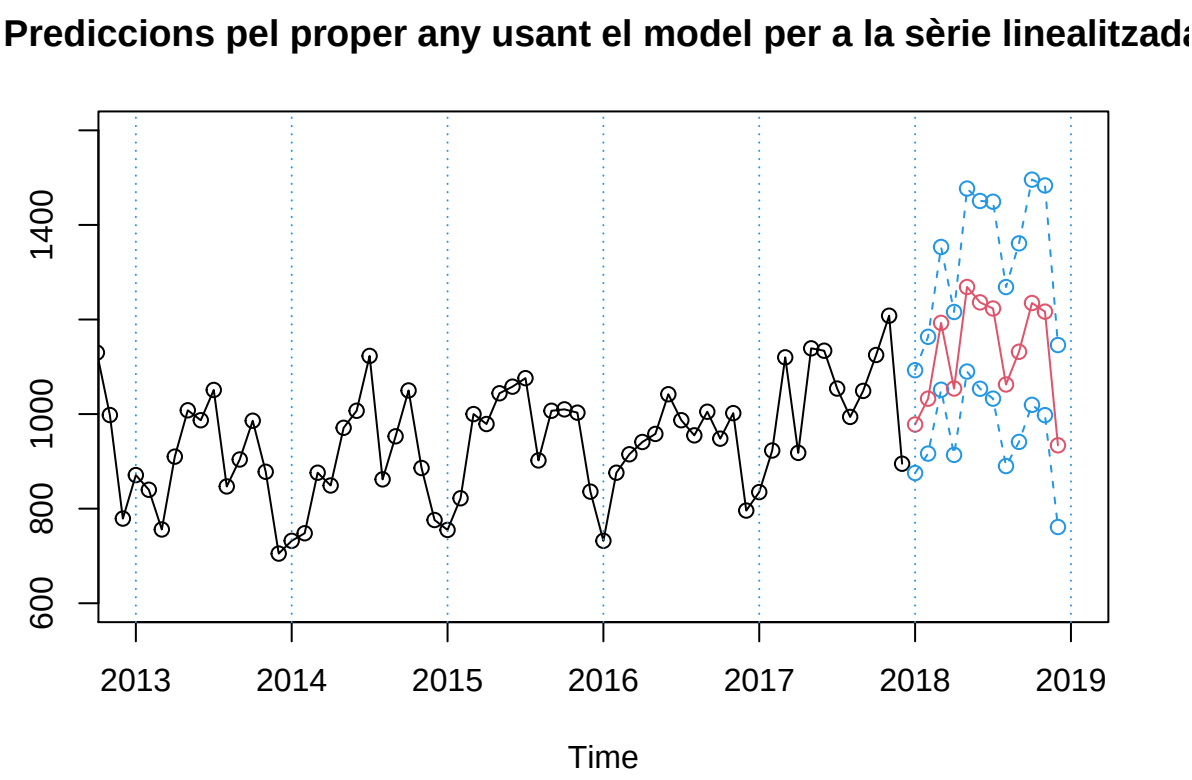
Veiem que, al igual que passava al primer model de la sèrie original, l'interval de predicció no s'ajusta a la sèrie original. Això és degut a que la variabilitat de les dades no ve explicada només pel temps i l'interval de confiança és bastant més ajustat que abans. Passarem a comparar els resultats d'aquestes prediccions amb els obtinguts amb el primer model.

```
## [1] 0.03323939
```

```
## [1] 0.03481084
```

Observem que hem obtingut més error amb aquestes prediccions utilitzant el model de la sèrie linealitzada que amb el de la sèrie original. Això és degut a que la variància de la sèrie linealitzada transformada era menor que la de l'original i, en conseqüència, els intervals de predicció són més ajustats, de manera que si la sèrie es comporta d'una manera poc esperada degut a la volabilitat de les dades, l'error serà més gran a aquest útlim model que hem plantejat.

Per acabar, farem la previsió per als propers 12 mesos de la sèrie original utilitzant el model per a la sèrie linealitzada.



Observem que ara l'interval de confiança s'ajusta més a la predicció i sembla prendre més en compte el comportament creixent dels darrers anys. Tot i així, després de veure que la volabilitat de les dades pot provocar un comportament poc esperat de la sèrie, no podem assegurar que realment aquesta predicció s'ajusti a la realitat.