# Práctica Series Temporales - Predicción

*Alvaro Ferro Perez*

*18/enero/2019*

## Resumen Ejecutivo

El objetivo de este informe es estimar modelos de series temporales sobre un conjunto de datos dado, tratando de predecir las cuotas de mercado de dos importantes empresas las cuales a primera vista parece que se comportan de manera opuesta. Es por ello que deberemos trabajar con variables temporales, siendo estas imprescindbles en nuestro modelo.

Para la predicción de nuestras cuotas desarrollaremos los conocidos modelos ARIMA, y por último incluiremos en nuestro análisis los modelos de transferencia que son una herramienta útil para evaluar impactos en las empresas y con ello reconducir los outliers, en nuestro caso, como veremos más adelante será la semana 135, que coincide con la primera de agosto, que fué cuando se produjo un acto en el que el Consejo de Terapéutica Dental de la American Dental Association (ADA) aprobó a Crest como una "ayuda importante en cualquier programa de higiene dental" lo que conllevó a un aumento de las ventas de Crest.

De todos los procesos estocásticos conocidos, tendremos en cuenta principalmente dos de ellos, ruido blanco, el cual es una sucesión de variables aleatorias con esperanza igual a cero, varianza constante e independiente para diferentes valores de t(covarianza nula).

El modelo ARIMA es un modelo autoregresivo, significa que si la variable endógena durante un periodo se puede explicar mediante sucesos pasados y añadiéndole un término del error. Si tiene una distribución normal, la teoría nos indica que bajo ciertas condiciones previas, toda la información la podemos expresar como una combinación lineal de sus valores pasados, para ello debemos asegurarnos que es una serie estacionaria y si no lo es, como es nuestro caso, debemos de transformar la serie original. Utilizaremos tanto el análisis gráfico como el econométrico para analizar la tendencia y la estacionaridad de los datos. Realizaremos la predicción sobre las ultimas 16 semanas de la empresa Crest y de Colgate.

Una de sus ventajas es que proporciona predicciones óptimas, y nos permite elegir entre un amplio rango de distintos modelos que represente el mejor comportamiento de los datos. Y tiene una serie de requisitos como el principio de parsimonia, el cual, es utilizado normalmente en matemáticas que lo que nos indica que es mejor utilizar un polinomio simple a diferencia de un polinomio complejo. Se exige que la serie temporal que estemos tratando sea estacionaria ya que eso permite ajustar mucho mejor la media y varianza, otros supuestos como el de ruido blanco.También hay que tener en cuenta la bondad del ajuste, es decir que el modelo se ajuste bien a los datos, y evidentemente que las predicciones sean correctas.

La formulacion de modelos arima permite incluir algunos de los modelos de alisado exponencial. Nuestro parámetro de media móvil 0 coincide con 1-alpha,siendo alpha el parámetro aislado. Por tanto, el objetivo de este informe será determinar si los efectos sobre la empresa 'Crest' influyen en 'Colgate' que, como veremos, solo lo hacen durante un corto periodo de tiempo.

## Análisis exploratorio de datos

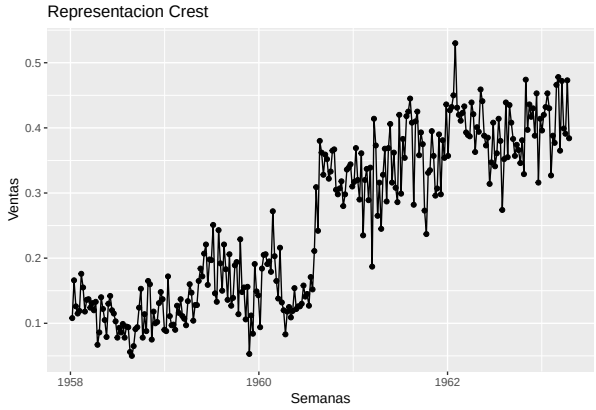
Nuestro trabajo comienza con el análisis y exploración de los datos aunque, en este caso concreto no son necesarios muchos cambios en las observaciones. Tenemos 276 observaciones y 4 variables, las variables son 'Crest' que corresponde a la cuota de mercado de dicha empresa al igual que 'Colgate', y las dos restantes son el año y la semana correspondiente a cada empresa.

Nuestra muestra abarcará todos nuestros datos dejando fuera las últimas 16 semanas, que son aquellas sobre las que queremos realizar la predicción de las cuotas de mercado de dichas empresas con el modelo ARIMA.

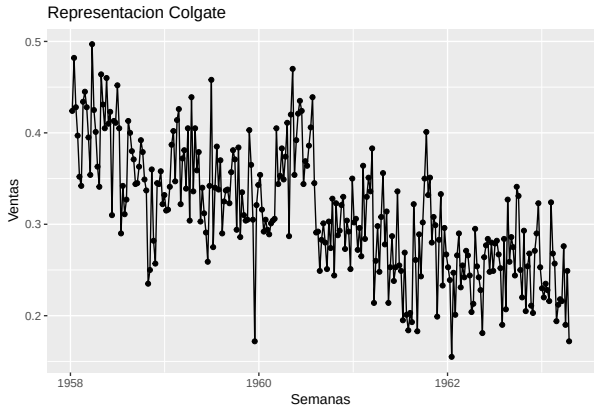
Es necesario primero hacer un parseo de las fechas para que adquieran el formato de *time-series* para, posteriormente poder realizar el *forecast* y que no haya ningún problema. Estamos trabajando en formato

semanal ya que es el que nos viene dado en los datos.

```
autoplot(zCuotaCrest) + geom_point() +
  ylab("Ventas")+ggtitle("Cuota semanal Crest")+xlab("Semanas") +
  ggtitle('Representacion Crest')
```




```
autoplot(zCuotaColgate) + geom_point() +
  ylab("Ventas") + ggtitle("Cuota semanal Colgate") + xlab("Semanas") +
  ggtitle('Representacion Colgate')
```




Como podemos observar en la gráfica los valores de 'Crest' aumentan constantemente, sin volver en ningún momento a un estado anterior, eso denota que estamos ante un 'escalón' o 'step' y no un 'impulso' o 'pulse', ya que las medias no vuelven a los valores anteriores.

En el caso de Colgate, este se comporta de manera similar al anterior solo que de manera bajista y de tipo 'escalón' también.

Nuestra seria temporal es no estacionaria en media porque tiene tendencia creciente en el caso de Crest y decreciente en el caso de Colgate. Estacionalidad tampoco presenta ya que la venta de este tipo de productos no se ve influida por la época del año en la que estemos como puede ser el caso de la luz o el agua. Para la implementación del modelo que queremos plantear podemos convetir la serie en estacionaria mediante logaritmos para hacer estacionaria la varianza o por diferenciación para la media por ejemplo.

## Modelo Arima

Ahora comenzaremos con el modelo ARIMA propiamente dicho, entrenaremos varios modelos autoarima para contrastar los resultados. La varianza, la hacemos estacionaria con el logaritmo, y la media mediante la diferencia. En el caso de la autocorrelación, que es la correlación de una variable consigo misma, si es

alta es algo bueno, eso quiere decir que podemos predecir la variable en función de ella misma. Más tarde buscaremos limpiarla de ruido, para hacerla estacionaria.

Se omiten del modelo las 16 semanas que comentamos al inicio de este informe, que pertenecen a las semanas sobre las cuales haremos nuestra predicción para ambas empresas.

```
fit1 = auto.arima(oVentasCrest)
fit2 = auto.arima(oVentasCrest, lambda = 0)
fit3 = auto.arima(oVentasCrest, lambda = 0, approximation = F, stepwise = F)
fit4 = auto.arima(oVentasCrest, ic = 'aic', trace = T)
```

```
##
##  Fitting models using approximations to speed things up...
##
##  ARIMA(2,1,2) with drift         : -852.9508
##  ARIMA(0,1,0) with drift         : -758.1344
##  ARIMA(1,1,0) with drift         : -825.2909
##  ARIMA(0,1,1) with drift         : -857.6885
##  ARIMA(0,1,0)                    : -760.0146
##  ARIMA(1,1,1) with drift         : -856.0304
##  ARIMA(0,1,2) with drift         : -856.6503
##  ARIMA(1,1,2) with drift         : -854.1478
##  ARIMA(0,1,1)                    : -858.4846
##  ARIMA(1,1,1)                    : -857.1803
##  ARIMA(0,1,2)                    : -857.619
##  ARIMA(1,1,2)                    : -855.2361
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(0,1,1)                    : -864.1502
##
##  Best model: ARIMA(0,1,1)
```

El modelo ARIMA, desde el punto de vista estocástico o moderno, tenemos trés parámetros de los que nos tenemos que preocupar, los cuales forman un modelo ARIMA no estacionario y se clasifica como un modelo "ARIMA (p, d, q)(P D Q)", donde la primera parte es la parte regular y la segunda sería la parte estacional la cual no tenemos en nuestro caso:

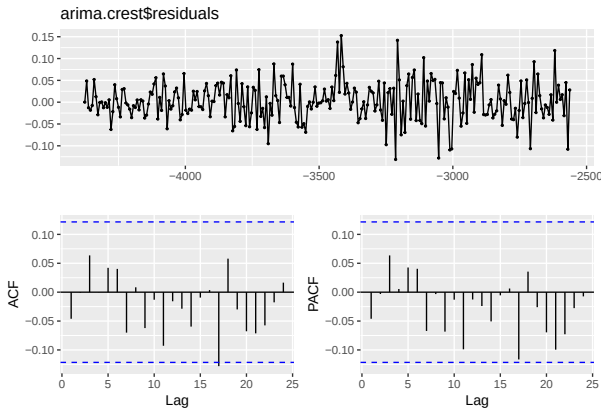
p es el número de términos autorregresivos, d es el número de diferencias no estacionales necesarias para la estacionalidad, y q es el número de errores de pronóstico retrasados en la ecuación de predicción.

La forma fácil de hallar un modelo adecuado sería realizar una comprobación de varios de ellos para finalmente quedarnos con aquel que menor AIC arroje. Así finalmente observamos que el mejor es el modelo ARIMA(0, 1, 1) con un AIC de -864.15 sin componente estacional.

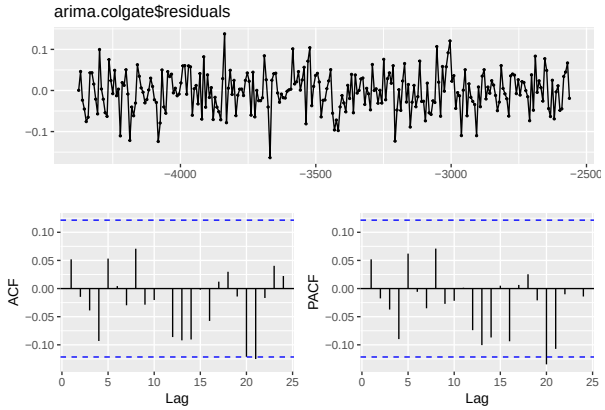
Es un modelo conocido como 'suavizado exponencial simple', en el cual es mejor, en vez de tomar la última media como único dato, tomar el promedio de las últimas observaciones para filtrar el ruido y estimar con mayor precisión la media local. Elegirá el que tenga el menor AIC.

El pronóstico de suavización exponencial simple es óptimo para patrones de demanda aleatorios o nivelados donde se pretende eliminar el impacto de los elementos irregulares históricos mediante un enfoque en períodos de demanda reciente.

```
ggtsdisplay(arima.crest$residuals)
```




```
ggtsdisplay(arima.colgate$residuals)
```




Como muestra serie temporal no es estacionaria, lo que tenemos que hacer es convertirla en estacionaria, mediante la diferenciación de orden D, una buena estrategia es comparar los ACF,que son los correlogramas de la función de autocorrelación. Como podemos observar en ambas, todos los datos se encuentran dentro de las bandas azules, eso nos indica que son ruido blanco y por tanto podemos continuar con el análisis. Ahora realizaremos el Text Box-Ljung, tanto con 'Colgate' como con 'Crest'.

```
Box.test(arima.crest$residuals,lag = 3, fitdf = 1, type = "Lj")
```

```
##
##  Box-Ljung test
##
## data:  arima.crest$residuals
## X-squared = 1.6314, df = 2, p-value = 0.4423
```

```
Box.test(arima.colgate$residuals,lag = 3, fitdf = 1, type = "Lj")
```

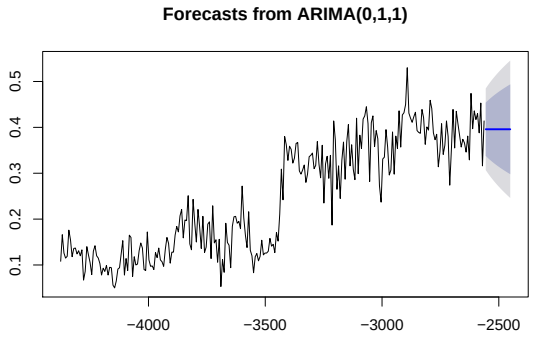
```
##
##  Box-Ljung test
##
## data:  arima.colgate$residuals
## X-squared = 1.1657, df = 2, p-value = 0.5583
```

Los gráficos de distribución de los errores hallados anteriormente tan solo nos dan un punto de vista primario, para asegurarnos del cumplimiento de la hipótesis de independencia de los residuos tendremos que usar el test de *Box-Ljung*

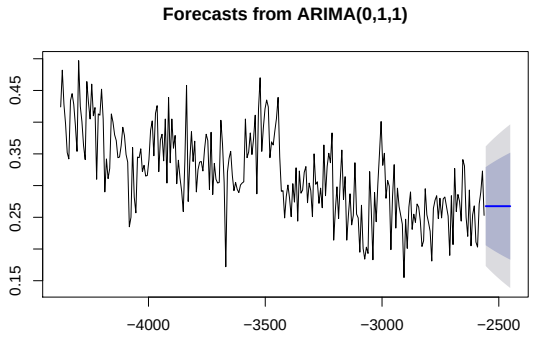
Este test lo que nos indica es como se distribuyen los residuos de los datos, es un contraste de hipótesis en el

que la hipótesis nula indica que los residuos de los datos se distribuyen de manera independiente, por tanto, eso querría decir que no existe autocorrelación entre los residuos y por tanto existe ruido blanco. Por tanto, buscamos un valor alto para nuestro P-valor con objetivo es aceptar la hipótesis nula, y eso nos indica que los residuos no tiene autocorrelación, gracias a esto podemos continuar con el análisis.

```
fventas.crest = forecast(arima.crest, h = 16)
plot(fventas.crest)
```




```
fventas.colgate = forecast(arima.colgate, h = 16)
plot(fventas.colgate)
```




Estos gráficos muestran la predicción sobre las 16 semanas para ambas empresas. Aunque esta predicción parezca algo rara, se puede intuir que la linea sigue la tendencia de los periodos anteriores de las empresas

Ahora vamos a proceder analizar los outliers tanto aditivos(afectan a la serie temporal) como innovativos(afectan al error), entonces vamos a analizar, los outliers para ambas empresas.

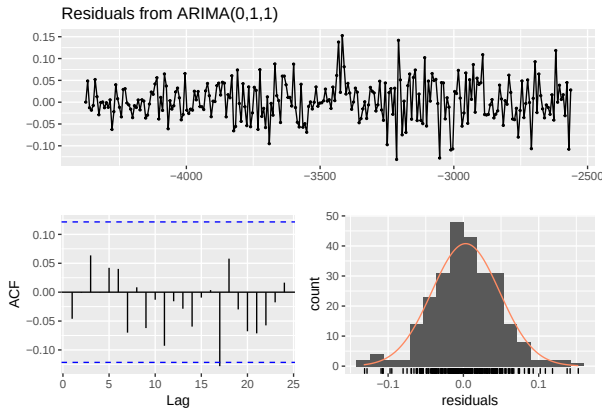
```
detectAO(arima.crest)
```

```
##               [,1]       [,2]       [,3]
## ind     135.000000 136.000000 138.000000
## lambda2   3.918954   4.372891   4.005427
```

```
detectIO(arima.crest)
```

```
## [1] "No IO detected"
```

```
checkresiduals(arima.crest)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)
## Q* = 4.9754, df = 9, p-value = 0.8364
##
## Model df: 1.   Total lags used: 10
```

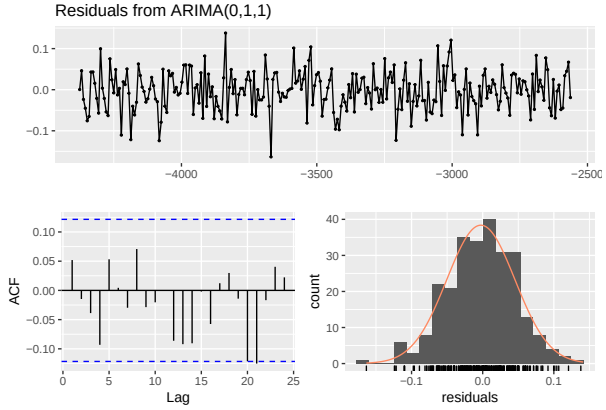
```
detectAO(arima.colgate)
```

```
## [1] "No AO detected"
```

```
detectIO(arima.colgate)
```

```
## [1] "No IO detected"
```

```
checkresiduals(arima.colgate)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)
## Q* = 6.1626, df = 9, p-value = 0.7235
##
## Model df: 1.   Total lags used: 10
```

Primeramente hemos de tener en cuenta que la semana 135 es la semana donde se produjo la crecida de Crest debido a un efecto externo. Puede ser normal que aparezca este valor como outlier. Además tenemos otros outliers en 136 y 138 para Crest.

No se encuentran errores innovativos en ninguna de las dos empresas.

```
coeftest(crest.arimax)
```

```
##
## z test of coefficients:
##
##              Estimate Std. Error  z value  Pr(>|z|)
## ma1         -0.744474   0.049265 -15.1117 < 2.2e-16 ***
## error136     0.022461   0.043176   0.5202   0.60290
## error138     0.076833   0.041428   1.8546   0.06365 .
## primero-MA0  0.133593   0.032689   4.0868 4.373e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(colgate.arimax)
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error  z value  Pr(>|z|)
## ma1        -0.804825   0.043637 -18.4437 < 2.2e-16 ***
## first-MA0 -0.101544   0.027859  -3.6449 0.0002675 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como observamos en el test de coeficientes, en ambos casos, la observación 135, anteriormente mostrada por la detección de Outliers Aditivos, tiene mucha significatividad. Por tanto será este el valor de corte en el modelo de intervención que realizaremos a continuación.

En el caso de 136 y 138, aunque la detección de errores nos los haya mostrado, al pasarlos por el modelo de transferencia no son significativos.

```
mod0 <- arimax(colgate_134_D,
               order=c(0,1,1),
               include.mean=TRUE,
               xtransf=crest_134_D,
               transfer=list(c(0,15)),
               method="ML")

coeftest(mod0)
```

```
##
## z test of coefficients:
##
##            Estimate Std. Error  z value  Pr(>|z|)
## ma1      -0.9999995  0.0214756 -46.5645 < 2.2e-16 ***
## T1-MA0   -0.5329365  0.1553428  -3.4307  0.000602 ***
## T1-MA1    0.0075869  0.1852828   0.0409  0.967338
## T1-MA2   -0.0431039  0.2016641  -0.2137  0.830749
## T1-MA3    0.1526312  0.2077696   0.7346  0.462573
## T1-MA4    0.0105404  0.2072262   0.0509  0.959434
## T1-MA5   -0.1105735  0.2041136  -0.5417  0.588008
## T1-MA6    0.0390328  0.2036775   0.1916  0.848024
## T1-MA7   -0.1783823  0.2003697  -0.8903  0.373323
## T1-MA8    0.0547611  0.2004141   0.2732  0.784669
## T1-MA9   -0.1667349  0.2022635  -0.8243  0.409743
```

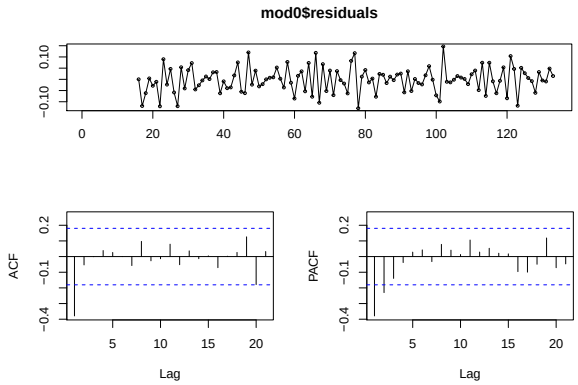
```
## T1-MA10  0.0749276  0.2031027   0.3689  0.712191
## T1-MA11  0.2016599  0.2061440   0.9782  0.327952
## T1-MA12  0.0762681  0.2069359   0.3686  0.712456
## T1-MA13  0.0630490  0.2003109   0.3148  0.752947
## T1-MA14 -0.1262427  0.1821569  -0.6930  0.488282
## T1-MA15 -0.0965627  0.1531617  -0.6305  0.528392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Conclusiones

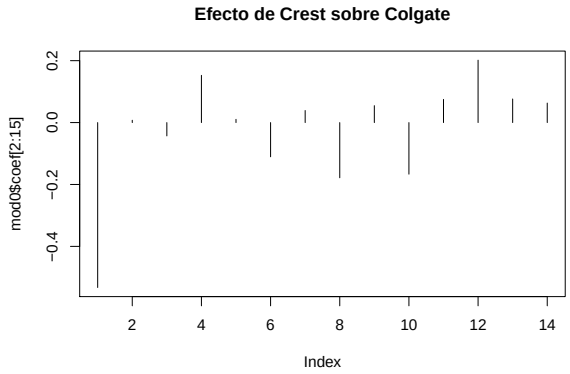
Hemos elegido el corte en 134 ya que este es la semana previa a efecto externo experimentado y además hemos convertido la serie para poder comparar ambas empresas de manera gráfica para ver de qué manera un efecto positivo sobre una ha hecho mella en la otra y si es de manera puntual o constante durante el tiempo a partir de ese valor.

Vemos que solo los primeros dos coeficientes aportan información al modelo, será con esos con los que continuemos.

```
tsdisplay(mod0$residuals)
```




```
plot(mod0$coef[2:15], type = 'h', main = "Efecto de Crest sobre Colgate")
```




```
mod <- arimax(colgate_134_D,
              order=c(0,1,1),
              include.mean=TRUE,
              fixed=c(NA,NA,0,0,NA),
              xtransf=crest_134_D,
```

```
              transfer=list(c(1,2)),
              method="ML")
```

Como podemos observar en el gráfico de la repercusión de Crest sobre Colgate, solo en el primer periodo de la serie se ha producido una caida muy importante dentro de las ventas de Colgate que coincide con la medida que se realizó la primera semana de Agosto, por tanto este aumento en la cuota de mercado de Crest se traduce en una caida dentro de la cuota de Colgate, por tanto se puede concluir que ambas empresas se influyen entre si. Es importante recalcar que el efecto solo es puntual en esa semana, ya que la empresa se volvió a estabilizar en las semanas posteriores.