Práctica final

Luis Llera García

17 de enero de 2019

EXECUTIVE SUMMARY

El problema de la predicción es una problema existente extrapolable a todos los campos de la economía que generalmente exige una serie de técnicas tanto estadísticas como econométricas muy complejas. Por tanto, en este informe trataremos de estimar y diagnosticar modelos dinámicos de series temporales en los que la variable tiempo juega un papel fundamental. Dentro de los modelos utilizados para predecir dicha variable se desarrollarán los conocidos modelos univariables ARIMA, y por último complementaremos nuestro análisis incorporando al estudio los conocidos modelos de transferencia, los cuales, son una herramienta que puede ser útil para evaluar impactos en las empresas y con ello reconducir los outliers representativos, en nuestro caso, como veremos más adelante será el outlier 135, que coincide con la primera semana 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 y las mismas no volvieron nunca al estado original ya que previamente antes del escalon se estaba produciendo un aumento progresivo de los datos, por lo tanto identificamos esta variación como un 'Step' no como un 'Impulso'. 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). La palabra ARIMA son las siglas Modelos Autorregresivos Integrados de Medias Móviles. 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 Ysubt la podemos expresarla como una combinación lineal de sus valores pasados, debemos asegurarnos que es una serie estacionaria y si no lo es 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 proporcionar 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. Antes de realizar el modelo Arima, tendremos que realizar el tratamiento y limpieza o depuración de los datos, que en este caso no nos ha llevado demasiado tiempo simplemente hemos tenido que generar una secuencia de fechas y pasar a formato zoo los datos que tenemos. Tanto la depuración de los datos como el modelo ARIMA los hemos realizado con el programa R-Studio, interfaz de R.

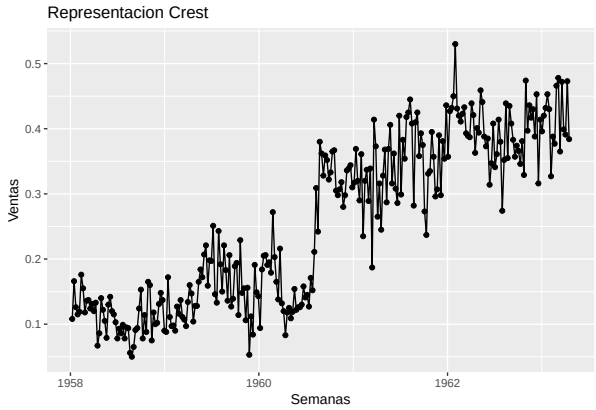
La formulacion de modelos ARIMA permite incluir algunos de los modelos de alisado esponencial, una de las equivalencias más importantes son las de un alisado explonencial simple, del que hablaremos más adelante. Nuestro parámetro de media movil 0 coincide con 1-alfa,siendo alfa el parámetro aislado. Por tanto, el objetivo de este informe será determinar si los efectos sobre la empresa 'Crest' influyen en 'Colgate'.

ANALISIS EXPLORATORIO DE DATOS

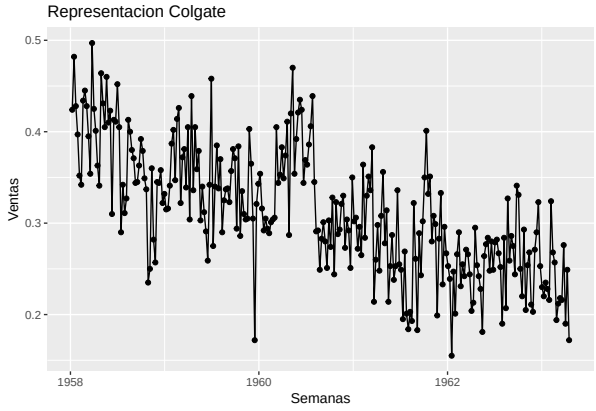
Como científico de datos vamos a comenzar el análisis con la limpieza y la depuración de los datos. Nuestros datos se componen de 276 observaciones y 4 variables, las variables son 'Crest' que corresponde a la cuota de mercado de dicha empresa al igual que 'Colgate', y las dos restantes son el año y 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. Primeramente, debemos de representar ambas empresas a lo largo de los años: Como podemos observar en el caso de crest, tiene una clara tendencia alcista a lo largo de toda la serie pero especialmente desde el verano del año 1960, la estacionalidad brilla por su ausencia, en una semana pasa de tener una cuota de mercado del 0.211 al 0.309, que intuitivamente coincide con la primera semana de agosto que fue cuando se produjo el acto comentado en la introducción.

```
#Primera aproximacion
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 su estado inicial, eso denota que estamos ante un 'escalón' o 'step' y no un 'impulso' o 'pulse', ya que las medias no vuelven a los valores iniciales. Y muestra una tendencia alcista notoria y el escalón en 1960 debido al acto que tuvo lugar ahí, pero pese a eso ya tenía una tendencia mas o menos alcista, por eso cabe suponer que 'Crest' se afianzó en el mercado de dentífricos a partir de los años 60.

Como podemos observar Colgate tiene prácticamente la misma representación que 'Crest' pero en sentido inverso muestra una tendencia bajista y aparentemente parece que no tiene estacionalidad. 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.

La función polinómica tambíen la tendremos que tener en cuenta, será en lo primero que nos tendremos que preocupar, en nuestro caso, tenemos una función polinómica, la cual, depende de los siguientes parámetros(b=1,s=0,r=0), esto nos indica que nos encontramos ante un escalón, y otra serie de indicaciones de las

que hablaremos en profundidad más adelante.

Como se puede observar tambien en este gráfico a parte de la ausencia de estacionariedad, podemos decir que tampoco tiene estacionalidad por que la cuota de mercado no se ve afectada por el mes en el que nos encontremos. Nuestra serie temporal es no estacionaria en media cuando tiene tendencia creciente o decreciente o cambios de nivel.

MODELO ARIMA

Ahora comenzaremos con el modelo ARIMA propiamente dicho, entrenaremos varios modelos autoarima para contrastar los resultados, en lineas generales un modelo es estacionario, en media,varianza y autocorrelación constante. La varianza, la hacemos estacionaria con el logaritmo, y la media mediante la diferencia y 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 limpiar los errores de ruido.

```
#ARIMA MODEL
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)
```

```
summary(fit4)
```

```
## Series: oVentasCrest
## ARIMA(0,1,1)
##
## Coefficients:
##          ma1
##       -0.6494
## s.e.   0.0448
##
## sigma^2 estimated as 0.002054:  log likelihood=434.08
## AIC=-864.15   AICc=-864.1   BIC=-857.04
```

```
##
## Training set error measures:
##                       ME       RMSE        MAE       MPE     MAPE
## Training set 0.003018071 0.04514444 0.03445351 -2.924537 17.27068
##                    MASE        ACF1
## Training set 0.1405405 -0.04605961
```

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)",que es la parte regular, y el segundo parámetro sería el estacional pero como no lo tenemos podemos deducir que no lo es no tiene este parámetro ARIMA(P,D,Q)s.

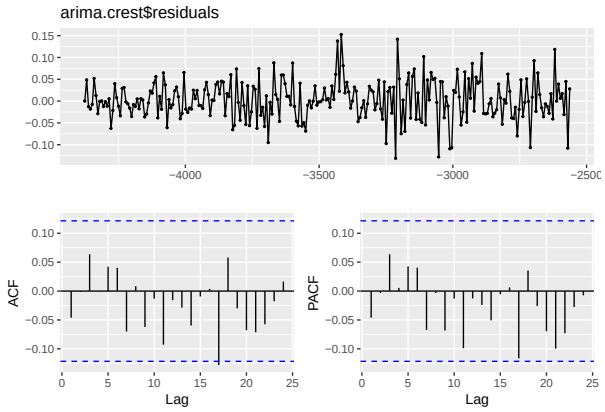
A continuación definiremos los diferentes parámetros:

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

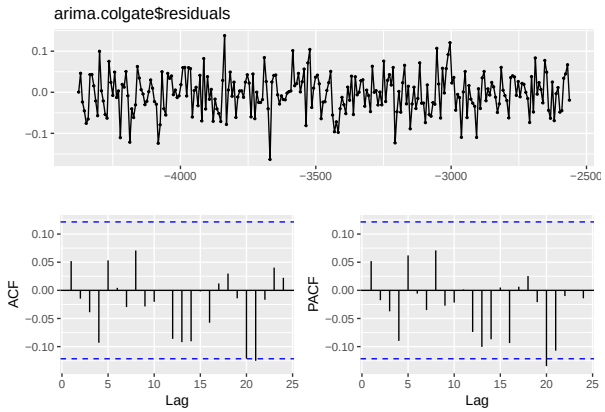
Como podemos observar, en las diferentes pruebas nos arrojan los mismos resultados, que el mejor modelo ARIMA es el (0,1,1), teniendo en cuenta el parámetro del AIC, elegirá al menor de ellos, que en nuestro caso es -864.15.

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. 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 para lograr óptimos resultados.

```
#residual analysis
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 de la función de autocorrelación. Como podemos observar en ambas, todos los datos se encuentran dentro de las bandas azules, eso nos indica que los residuos 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")
```

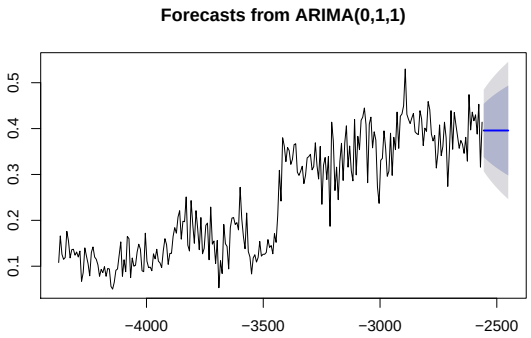
```
##
##  Box-Ljung test
##
## data:  arima.crest$residuals
## X-squared = 1.6314, df = 2, p-value = 0.4423
```

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

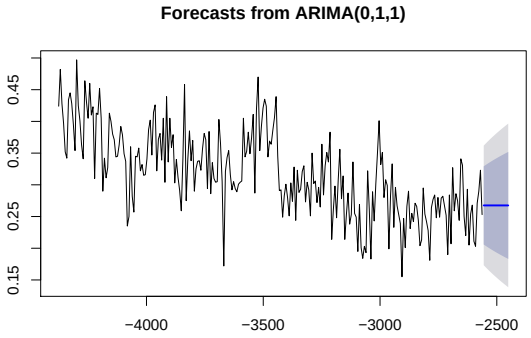
```
##
##  Box-Ljung test
##
## data:  arima.colgate$residuals
## X-squared = 1.1657, df = 2, p-value = 0.5583
```

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)
```



Como podemos observar, el forecast nos indica la predicción y podemos observar que tiene una predicción correcta ya que sigue la tendencia.

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

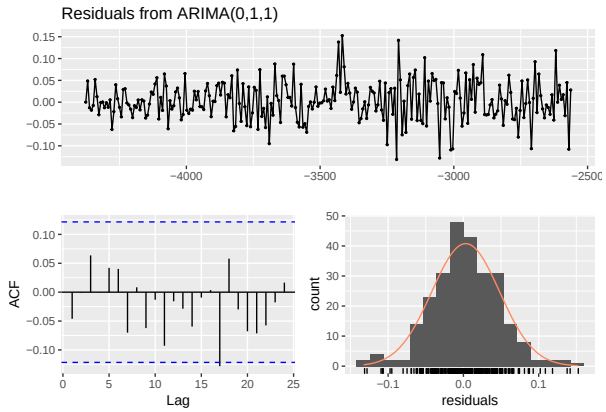
```
detectAO(arima.crest) #Outlier en 135/136/138
```

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

```
detectIO(arima.crest) #Nada
```

```
## [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
```

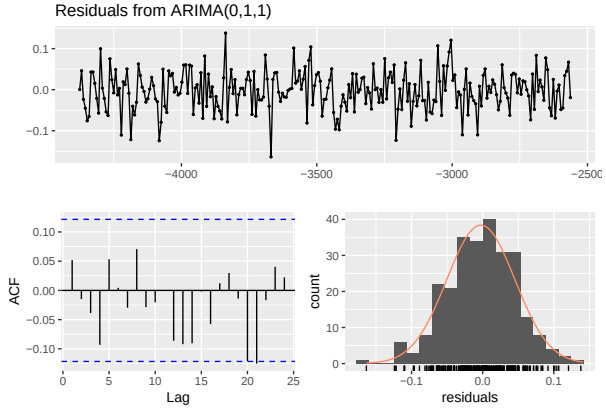
```
detectAO(arima.colgate)
```

```
## [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
```

En Crest obtenemos tres errores aditivos, a diferencia del de colgate, y con el gráfico podemos observar como los errores se distribuyen como una normal, en la semana 135 se encuentra incluida en los errores ya que fue cuando se produjo el acto que aumentó las ventas de 'Crest'.

```
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
```

Ahora realizamos el test de los coeficientes, en ambos casos la observación 135, que hemos mostrado anteriormente por la detección de outliers aditivos, tiene una significatividad alta, y por tanto será este el valor de corte en el modelo de intervención. De los dos restantes podemos prescindir.

CONCLUSIONES

```
mod0 <- arimax(colgate_134_D,
               order=c(0,1,1),
               include.mean=TRUE,
               xtransf=crest_134_D,
               transfer=list(c(0,15)), #funcion de transferencia con orden 15 numerador
               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
```

Estableceremos el corte en 134 debido a que coincide con la semana anterior al efecto comentado anteriormente a favor de Crest y además hemos convertido la serie para poder comparar las dos empresas, de manera gráfica, este efecto positivo afecta a Colgate y si es de manera constante durante el tiempo a partir de ese valor.

Los únicos coeficientes que aportan información son el primero y el segundo, debido a ello el análisis que realizaremos a continuación se basará en estos dos coeficientes.

Como podemos observar en el gráfico de 'Efecto de Crest sobre Colgate', lo que podemos observar es que en el primer periodo de la serie se ha producido una caida muy importante dentro de las ventas de Colgate que coincide perfectamente con la medida que se realizó la primera semana de Agosto, por tanto un aumento brutal en la couta de mercado de Crest se traduce en una caida brutal dentro de la cuota de mercado de Colgate, por tanto se puede concluir que ambas empresas se influyen entre si, pero solo durante ese periodo, después al ser un escalón nos demuestra que no vuelve a la situación inicial.