Documento discusión datos COVID-19

F.J. Alonso [a],

[a] *Department of Statistics and O.R., University of Granada, Spain*

Abstract

Este trabajo simplemente intenta presentar algunas vías para el tratamiento de los datos de COVID-19 usando técnicas simples de regresión y de análisis de series.

Key words: Regresión, predicción, series temporales, modelos ARIMA, modelos VAR (vectorial autorregresive).

1 Introducción

Los datos que voy a analizar son los proporcioinados por el ministerio y descargados de la web en el siguiente enlace. Esta redacción la estoy llevando a cabo antes de que se presente el dato del 31/marzo, aunque actualizaré mis conclusiones al final.

`https://datos.gob.es/es/catalogo/`
`e05070101-evolucion-de-enfermedad-por-el-coronavirus-covid-19`

El fichero descargado consta de 7 columnas de datos correspondientes a "código ISO de comunidad autonómica", "Fecha", "Casos", "Hospitalizados", "UCI", "Fallecidos", "Recuperados".

En una primera etapa me centraré en el estudio de la evolución de casos de COVID-19, y posteriormente intentaré una modelización conjunta de la situación hospitalaria (Hospitalizados, UCI) junto con los casos presentados.

* Corresponding author. Department of Statistics and Operations Research, University of Granada, Campus de Fuente Nueva s/n, E-18071 Granada, Spain. Phone: (34) 958243154.

Email address: `falonso@ugr.es` (F.J. Alonso).

Este documento, más que una propuesta de metodología, pretende presentar las técnicas que pueden llevar a un debate y la aplicación de nuevas herramientas u otras complementarias. He de reconocer que la "bondad" de los datos es muy pobre, ya que se han obtenido por procedimientos y protocolos que han ido modificándose en el tiempo y aún así, no se tiene la certeza, más bien todo lo contrario, de que el dato refleje el total de infectados (centrándose en "Casos"). También he de reconocer que las técnicas de series usuales no se deben de aplicar, ya que el número de datos es muy reducido. Box y Jenkins sugieren que al menos se debe de disponer de 50 datos para una modelización ARIMA univariante. En esta situación estamos muy lejos de disponer de ese volumen.

En este sentido, me salto todas las consideraciones anteriores, y paso a la descripción de lo que se podría hacer.

2 Cuando se alcanzará el máximo de casos (predicción lineal)

En primer lugar, voy a trabajar con los datos agrupados para toda la nación `Dat_ESP`. Voy a usar R, e iré presentando las partes del código que considere más interesantes, sobre todo para los que no estéis familiarizados con el manejo de series temporales.

Los datos agrupados aparencen en la figura 1

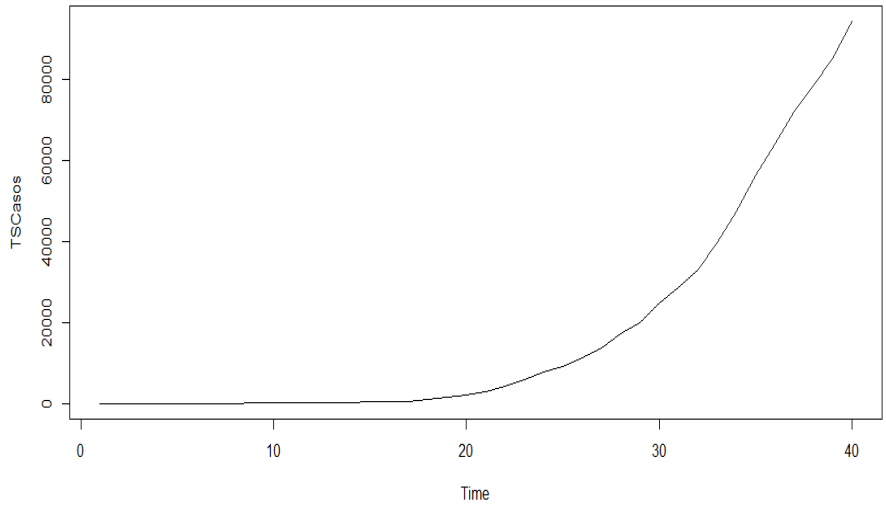

Fig. 1. Casos de COVID-19 hasta 30/03.

```
TSCasos<-ts(Dat_ESP$Casos)
```

```
TSCasos
 [1]     2     2     2     2     3    10    16    29    43
[10]    65   113   148   194   238   364   386   526  1006
[19]  1606  2117  2929  4130  5844  7698  9149 11178 13716
[28] 17147 19980 24926 28572 33089 39673 47610 56188 64059
[37] 72248 78797 85195 94417
```

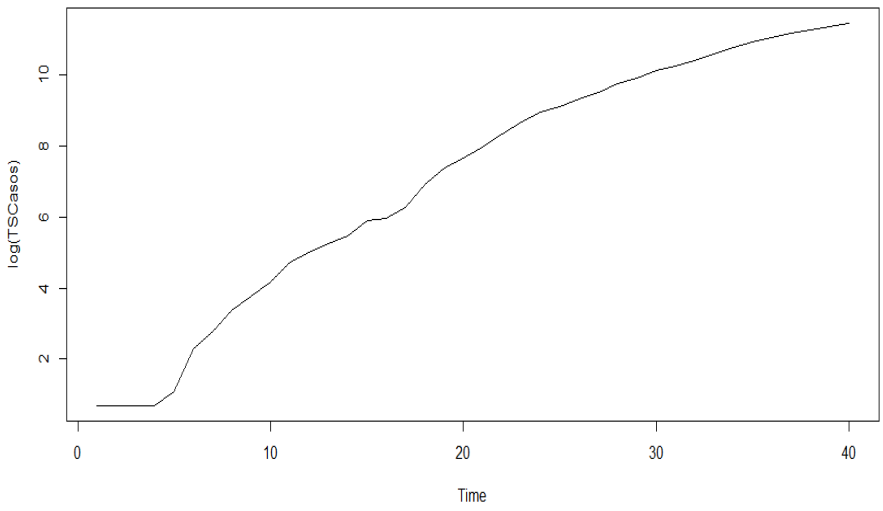

Fig. 2. Logaritmo de los casos de COVID-19 hasta 30/03.

Si le aplicamos logaritmos a los datos (logaritmos naturales) la gráfica que obtenemos se transforma en la figura 2

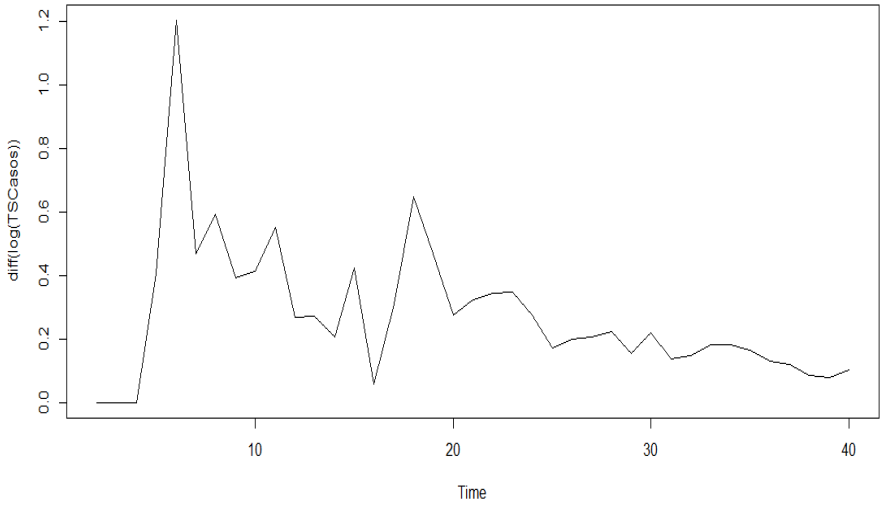

Fig. 3. Logaritmo de los casos de COVID-19 hasta 30/03.

Mi objetivo inicial fue intentar buscar el máximo de esa función, es decir, en qué punto temporal se alcanzaría la tan ansiada meseta. Para ello busco el punto en el que el incremento en la curva anterior sea cero, es decir realizo la diferencia $\log(Caso)_t - \log(Caso)_{t-1}$. Este gráfico aparece en la figura 3

Creo que podría ser una buena estimación de esa llegada, realizar una proyección de los valores finales y determinar el tiempo en el que estas diferencias toman el valor 0. En este sentido, empecé a "jugar" el domingo con los datos, y las predicciones eran muy alentadores, pero los datos del lunes y sobre todo el del martes ha corregido al alza este horizonte.

Para ello, realizo una regresión lineal simple, usando los últimos valores, desde el dato 25. El gráfico se puede ver en la figura 4. Con esta regresión el máximo (valor de incremento igual a 0) se alcanza el 44.36, es decir el 5 de abril (proyectando estos datos, en la rueda de prensa del 7 de abril se espera que no se produzca un incremento en el número de casos).

```
DLcasos<-diff(log(TSCasos), differences=1)
NCDL<-length(DLcasos)
model1<-lm(c(25:NCDL)~DLcasos[25:NCDL])
plot(DLcasos[25:NCDL],c(25:NCDL),xlim = c(0,0.24),ylim = c(25,43))
abline(model1)
```

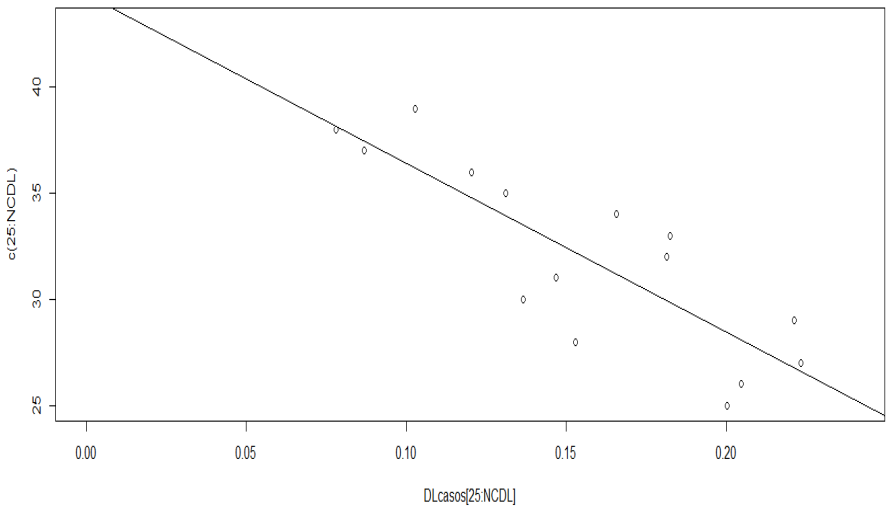

Fig. 4. Regresión realizada el martes 30.

Si presentamos los resultados realizados el domingo (figura 5) y el lunes (figura 6) que corresponden a quitar dos y un dato al final de la serie respectivamente, obtendríamos que ese máximo se alcanzaría con los datos disponibles el domingo el 42.85 (3 de abril) y con los datos disponibles el lunes el 43.25 (4 de abril). Las tres líneas de predicción juntas se pueden apreciar en el gráfico

7, donde la línea roja corresponde al domingo, la azul al lunes y la negra el martes.

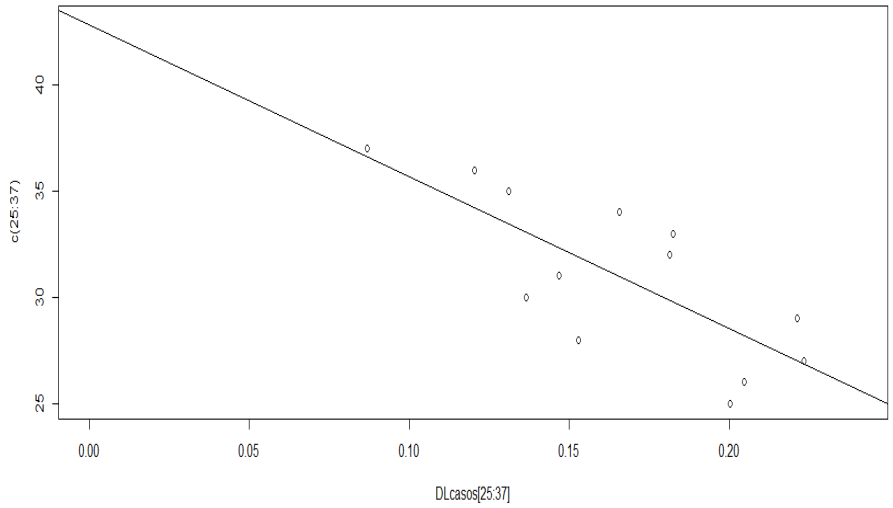

Fig. 5. Regresión realizada el domingo 28.

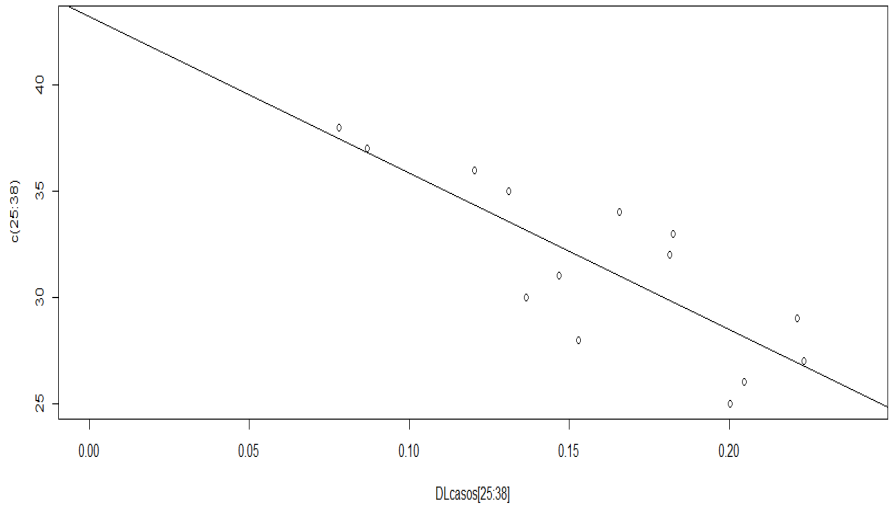

Fig. 6. Regresión realizada el lunes 29.

Personalmente creo que el ajuste lineal no va a ser muy adecuado para la predicción de este horizonte, ya que parece, como indican los datos de Italia, aunque su dato de ayer fue muy bueno, que al final, el decrecimiento va a ser más lento que un lineal. Sin embargo pienso que este abordaje del problema puede suscitar discusión y un nuevo enfoque para dar aproximaciones rápidas y dinámicas.

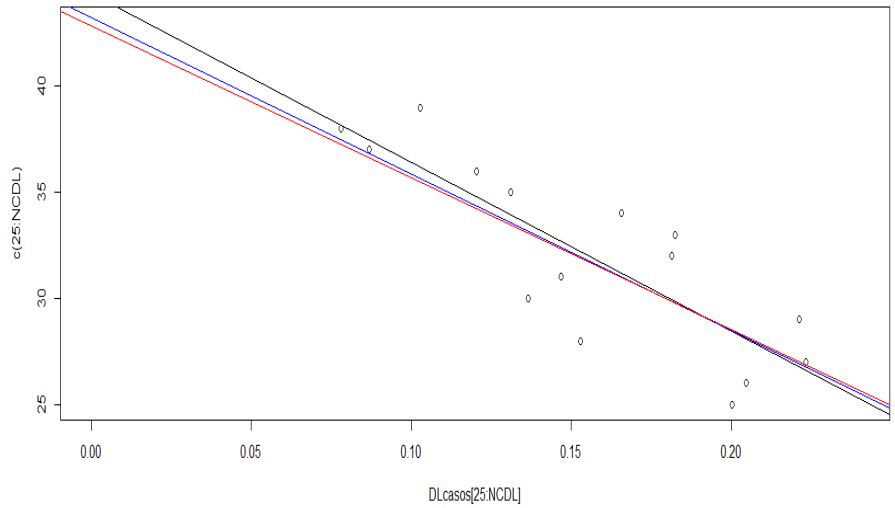

Fig. 7. Regresión con los tres conjuntos de datos (rojo=domingo, azul=lunes, negro=martes).

3 Predicción usando modelos se series multivariantes

He de decir, que como en el caso anterior, voy a trabajar con las series de logaritmos diferenciadas $\log(X)_t - \log(X)_{t-1}$, tanto para los casos, como para los valores de hospitalizados y de ingresos en la UCI. Mucho más cuidado hay que tener en esta situación ya que empieza a haber datos de hospitalizados el 10 de marzo, aunque había ingresos en la UCI desde dos días antes. Así los datos temporales completos de los que se dispone es más reducido aún.

Si se realiza una correlación cruzada, para ver la influencia entre las variables (más bien entre los residuos de una modelización univariante), la única influencia que se detecta es entre los casos y los hospitalizados, con una demora de tres días (ver figura 8).

Esta influencia, también la podemos estudiar usando el test de causalidad de Granger que es significativo, es decir la hospitalización se ve influenciada por el número de casos aparecidos.

```
grangertest(DLcasos,DLHosp,order=3)
Granger causality test

Model 1: DLHosp ~ Lags(DLHosp, 1:3) + Lags(DLcasos, 1:3)
Model 2: DLHosp ~ Lags(DLHosp, 1:3)
  Res.Df Df      F   Pr(>F)
1     10
```

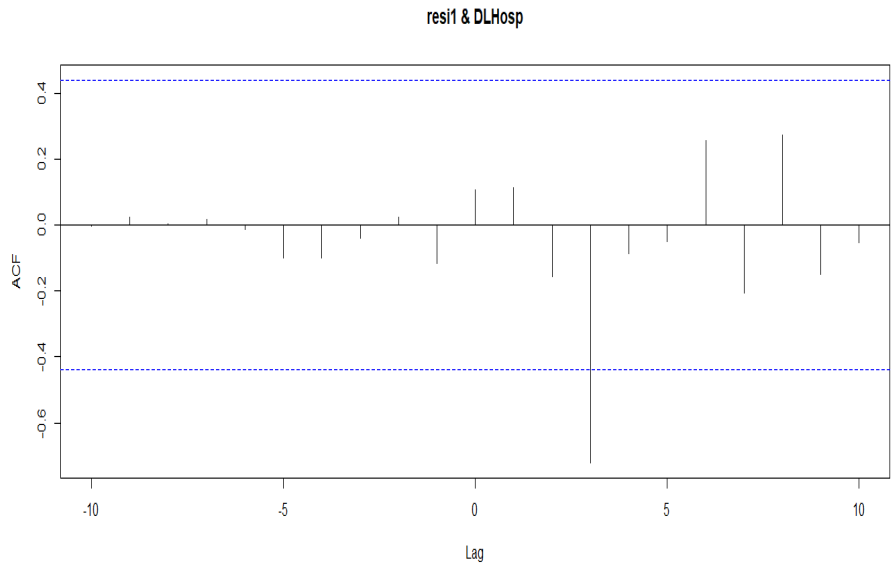

Fig. 8. Correlación cruzada (ccf) entre la serie 20 últimos residuos ARIMA(0,1,1) DLCasos y DLHosp.

```
2     13 -3 7.3537 0.006872 **
```

También los casos en la UCI se ven influenciados, según el test de Granger, por el número de hospitalizados

```
Granger causality test

Model 1: DLUCI ~ Lags(DLUCI, 1:3) + Lags(DLHosp, 1:3)
Model 2: DLUCI ~ Lags(DLUCI, 1:3)
  Res.Df Df      F  Pr(>F)
1     10
2     13 -3 3.9071 0.04389 *
```

En el otro sentido, y en ninguna de las otras dos variaciones, no se podría rechazar, al 5%, la ausencia de causalidad.

Para realizar una modelización rápida, uso un modelo VAR realizando una regresión sobre los valores temporales (t-1), (t-2) y (t-3), considerando también un término de tendencia determinística lineal.

$$\begin{pmatrix} \boldsymbol{DC} \\ \boldsymbol{DH} \\ \boldsymbol{DUCI} \end{pmatrix}_t = \begin{pmatrix} a + b \times t \\ a' + b' \times t \\ a'' + b'' \times t \end{pmatrix} + A_1 \begin{pmatrix} \boldsymbol{DC} \\ \boldsymbol{DH} \\ \boldsymbol{DUCI} \end{pmatrix}_{t-1} + A_2 \begin{pmatrix} \boldsymbol{DC} \\ \boldsymbol{DH} \\ \boldsymbol{DUCI} \end{pmatrix}_{t-2} + A_3 \begin{pmatrix} \boldsymbol{DC} \\ \boldsymbol{DH} \\ \boldsymbol{DUCI} \end{pmatrix}_{t-3} + \varepsilon_t, \quad (1)$$

donde DC, DH y DUCI denotan los valores con los que estamos trabajando (diferencias de logaritmos), "aes" y "bes" son constantes y A_i son matrices 3×3.

```
library(vars)
multiN<-Dat_ESP[,c("Casos","Hospitalizados","UCI")][20:ndias,]
multiN<-diff(ts(log(multiN)))
EQ1<-VAR(multiN,p=3,type=c("both"),ic="AIC")
summary(EQ1)
```

La salida que se obtiene es

```
EQ1

VAR Estimation Results:
=======================

Estimated coefficients for equation Casos:
==========================================
Call:
Casos = Casos.l1 + Hospitalizados.l1 + UCI.l1 + Casos.l2 +
Hospitalizados.l2 + UCI.l2 + Casos.l3 + Hospitalizados.l3 + UCI.l3
+ const + trend

         Casos.l1 Hospitalizados.l1           UCI.l1
       0.06224567        0.03170036       -0.05912448
         Casos.l2 Hospitalizados.l2           UCI.l2
       0.18326092        0.04791914       -0.20654262
         Casos.l3 Hospitalizados.l3           UCI.l3
      -0.41763237       -0.07237529        0.04433970
            const             trend
       0.40763364       -0.01335363
```

```
Estimated coefficients for equation Hospitalizados:
===================================================
Call:
Hospitalizados = Casos.l1 + Hospitalizados.l1 + UCI.l1 + Casos.l2 +
Hospitalizados.l2 + UCI.l2 + Casos.l3 + Hospitalizados.l3 + UCI.l3
+ const + trend

         Casos.l1 Hospitalizados.l1            UCI.l1
       1.05836282       -0.66887857       -0.28279533
         Casos.l2 Hospitalizados.l2            UCI.l2
       0.13523354       -0.19932775        0.20535644
         Casos.l3 Hospitalizados.l3            UCI.l3
       0.24652136       -0.24236987        0.44898098
            const              trend
       0.30470713       -0.01767784


Estimated coefficients for equation UCI:
========================================
Call:
UCI = Casos.l1 + Hospitalizados.l1 + UCI.l1 + Casos.l2 +
Hospitalizados.l2 + UCI.l2 + Casos.l3 + Hospitalizados.l3 + UCI.l3
+ const + trend

         Casos.l1 Hospitalizados.l1            UCI.l1
      -1.20201126        0.73987750       -0.27013092
         Casos.l2 Hospitalizados.l2            UCI.l2
       0.58673983        0.12626617       -0.05176503
         Casos.l3 Hospitalizados.l3            UCI.l3
      -0.44041093       -0.19489549       -0.04676715
            const              trend
       0.56569699       -0.02085945

summary(EQ1)

VAR Estimation Results:
=======================
Endogenous variables: Casos, Hospitalizados, UCI
Deterministic variables: both
Sample size: 17
Log Likelihood: 139.528
Roots of the characteristic polynomial:
0.8503 0.8503 0.8411 0.8411 0.7604 0.7604 0.5567 0.4589 0.4589
Call:
VAR(y = multiN, p = 3, type = c("both"), ic = "AIC")
```

```
Estimation results for equation Casos:
======================================
Casos = Casos.l1 + Hospitalizados.l1 + UCI.l1 + Casos.l2 + Hospitalizados.l2 + UCI

                   Estimate Std. Error t value Pr(>|t|)
Casos.l1           0.062246   0.389654   0.160   0.8783
Hospitalizados.l1  0.031700   0.163152   0.194   0.8524
UCI.l1            -0.059124   0.157739  -0.375   0.7207
Casos.l2           0.183261   0.343591   0.533   0.6129
Hospitalizados.l2  0.047919   0.060943   0.786   0.4616
UCI.l2            -0.206543   0.166989  -1.237   0.2624
Casos.l3          -0.417632   0.297510  -1.404   0.2100
Hospitalizados.l3 -0.072375   0.050446  -1.435   0.2014
UCI.l3             0.044340   0.114191   0.388   0.7112
const              0.407634   0.198350   2.055   0.0856 .
trend             -0.013354   0.006686  -1.997   0.0928 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.02932 on 6 degrees of freedom
Multiple R-Squared: 0.8833,Adjusted R-squared: 0.6887
F-statistic:  4.54 on 10 and 6 DF,  p-value: 0.03874


Estimation results for equation Hospitalizados:
===============================================
Hospitalizados = Casos.l1 + Hospitalizados.l1 + UCI.l1 + Casos.l2 + Hospitalizados

                   Estimate Std. Error t value Pr(>|t|)
Casos.l1           1.058363   0.454114   2.331   0.0586 .
Hospitalizados.l1 -0.668879   0.190142  -3.518   0.0126 *
UCI.l1            -0.282795   0.183834  -1.538   0.1749
Casos.l2           0.135234   0.400431   0.338   0.7471
Hospitalizados.l2 -0.199328   0.071024  -2.806   0.0309 *
UCI.l2             0.205356   0.194614   1.055   0.3320
Casos.l3           0.246521   0.346727   0.711   0.5038
Hospitalizados.l3 -0.242370   0.058791  -4.123   0.0062 **
UCI.l3             0.448981   0.133082   3.374   0.0150 *
const              0.304707   0.231163   1.318   0.2355
trend             -0.017678   0.007792  -2.269   0.0638 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.03417 on 6 degrees of freedom
Multiple R-Squared: 0.961,Adjusted R-squared: 0.8961
F-statistic:  14.8 on 10 and 6 DF,  p-value: 0.001838


Estimation results for equation UCI:
====================================
UCI = Casos.l1 + Hospitalizados.l1 + UCI.l1 + Casos.l2 + Hospitalizados.l2 + UCI.l

                   Estimate Std. Error t value Pr(>|t|)
Casos.l1          -1.202011   0.357462  -3.363  0.01518 *
Hospitalizados.l1  0.739877   0.149672   4.943  0.00260 **
UCI.l1            -0.270131   0.144707  -1.867  0.11118
Casos.l2           0.586740   0.315205   1.861  0.11200
Hospitalizados.l2  0.126266   0.055908   2.258  0.06468 .
UCI.l2            -0.051765   0.153193  -0.338  0.74694
Casos.l3          -0.440411   0.272931  -1.614  0.15773
Hospitalizados.l3 -0.194895   0.046278  -4.211  0.00561 **
UCI.l3            -0.046767   0.104757  -0.446  0.67094
const              0.565697   0.181963   3.109  0.02088 *
trend             -0.020859   0.006133  -3.401  0.01448 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.02689 on 6 degrees of freedom
Multiple R-Squared: 0.9712,Adjusted R-squared: 0.9231
F-statistic: 20.22 on 10 and 6 DF,  p-value: 0.0007678



Covariance matrix of residuals:
                   Casos Hospitalizados        UCI
Casos          0.0008594      0.0002604  0.0002381
Hospitalizados 0.0002604      0.0011673 -0.0004942
UCI            0.0002381     -0.0004942  0.0007233

Correlation matrix of residuals:
               Casos Hospitalizados     UCI
Casos          1.000         0.2600  0.3020
Hospitalizados 0.260         1.0000 -0.5378
UCI            0.302        -0.5378  1.0000
```

La salida, cuando introducimos EQ1, indica las ecuaciones lineales que expresa cada variable unidimensional (DC, DH o DUCI) en función de la observación (los valores de a y b), dados por `const` y `trend`, y los coeficientes que multiplicarán a los valores retardados k tiempos. Así por ejemplo, 0.04791914 que es el valor del coeficiente que acompaña a `Hospitalizados.l2` en la primera ecuación `Casos=...`, indica que en la ecuación lineal, el correspondiente sumando sería $0.04791914 \times DH_{t-2}$.

Se debe de hacer un mayor trabajo con este tipo de modelos, eliminando variables que no sean significativas. Pero puede ser un primer acercamiento al problema y también puede generar debate y nuevas propuestas.

La predicción usando la ecuación multivariante estimada para un horizonte de predicción tres tiempos hacia adelante vendría dada por:

```
library(forecast)
forecast(EQ1,3)

Casos
   Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
22     0.09786835 0.06029792 0.1354388 0.04040934 0.1553274
23     0.09177797 0.05402964 0.1295263 0.03404688 0.1495091
24     0.07722087 0.03798469 0.1164571 0.01721431 0.1372274

Hospitalizados
   Point Forecast       Lo 80      Hi 80       Lo 95      Hi 95
22     0.03226090 -0.01152475 0.07604655 -0.03470347 0.09922528
23     0.03909750 -0.01838570 0.09658069 -0.04881546 0.12701046
24     0.01847149 -0.04018111 0.07712409 -0.07122992 0.10817290

UCI
   Point Forecast       Lo 80      Hi 80       Lo 95      Hi 95
22    0.023105525 -0.01136092 0.05757197 -0.02960636 0.07581741
23    0.019086521 -0.04557563 0.08374867 -0.07980570 0.11797874
24    0.000491897 -0.08612256 0.08710635 -0.13197348 0.13295727
```

El valor proporcionado hoy para el número de casos es 102.136, es decir resultaría 0,078584117. Menor que la predicción, pero entre los límites de confianza.

Hemos de recordar que si alcanzase el valor 0, estaríamos en una situación de haber llegado al máximo, a la "cima" como dicen los políticos. Aunque todavía nos movamos por valores de casos relativamente alto, es esperanzador que el número de hospitalizados e ingresos en la UCI prácticamente se estabiliza. Quizás, estos últimos datos sí sean más fiables y no se vean lastrados por las políticas en la toma de datos o el empleo o no de test sobre distintos grupos

de la población.

References

[1] Granger, C.W.J. (1969). "Investigating Causal Relations by Econometric Models and Cross Spectral Methods," Econometrica, 37, 424- 438.

[2] Hamilton, J.D. (1994). Time Series Analysis. Princeton University Press, Princeton.

[3] Lütkepohl, Helmut (2009). Econometric Analysis with Vector Autoregressive Models (in Handbook of Computational Econometrics). John Wiley & Sons, Ltd. New York. 281-319 doi:10.1002/9780470748916.ch8

[4] Tsay, R. (2001). Analysis of Financial Time Series. John Wiley & Sons. New York.

4 Actualización con los datos 31/03

Con el nuevo dato proporcionado hoy 1 de abril, la predicción obtenida es 45,046. Por lo tanto el máximo se alcanzaría aproximadamente en el día 5 de abril. La predicción se puede observar en la figura 9 (valor de la recta para x=0).

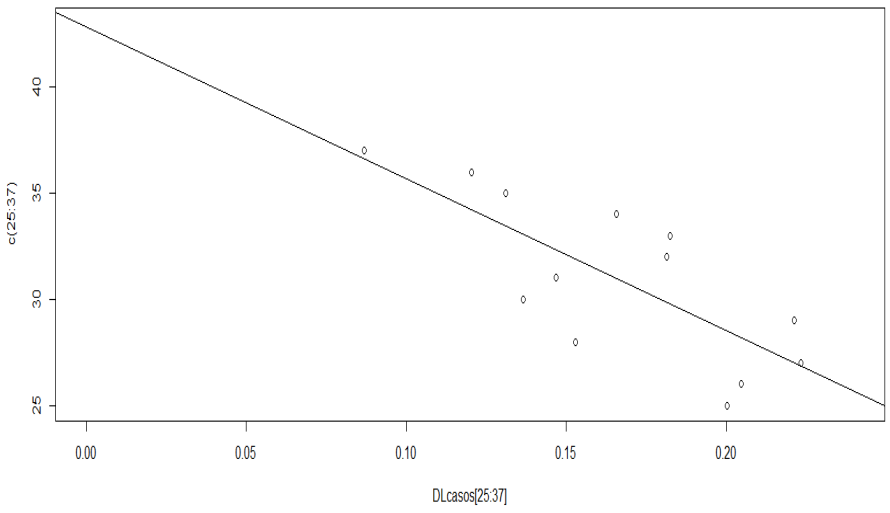

Fig. 9. Regresión realizada con los datos hasta 31/03.

Las predicciones usando el modelo multivariante con los datos actualizados son:

```
Casos
   Point Forecast      Lo 80      Hi 80        Lo 95     Hi 95
23     0.08327430 0.04775985 0.11878875 0.0289596403 0.1375890
24     0.05994474 0.02432012 0.09556935 0.0054615862 0.1144279
25     0.05728705 0.02003644 0.09453766 0.0003171601 0.1142569

Hospitalizados
   Point Forecast       Lo 80      Hi 80       Lo 95      Hi 95
23    0.008404487 -0.03233737 0.04914635 -0.05390481 0.07071378
24    0.024185991 -0.03114956 0.07952154 -0.06044243 0.10881441
25   -0.011570111 -0.06847530 0.04533508 -0.09859909 0.07545887

UCI
   Point Forecast       Lo 80      Hi 80        Lo 95     Hi 95
23     0.05166452  0.01862212 0.08470692  0.001130529 0.1021985
24    -0.02465478 -0.08494038 0.03563083 -0.116853649 0.0675441
25     0.02789081 -0.05451089 0.11029251 -0.098131715 0.1539133
```

Se produce una reducción en la "tasa de crecimiento" de número de casos con el nuevo dato, pasando a que mañana en vez de 0.09177797 se espera que se obtenga 0.08327430.