PAC 8 - Aproximació de funcions i regressió (II)

Amelia Martínez

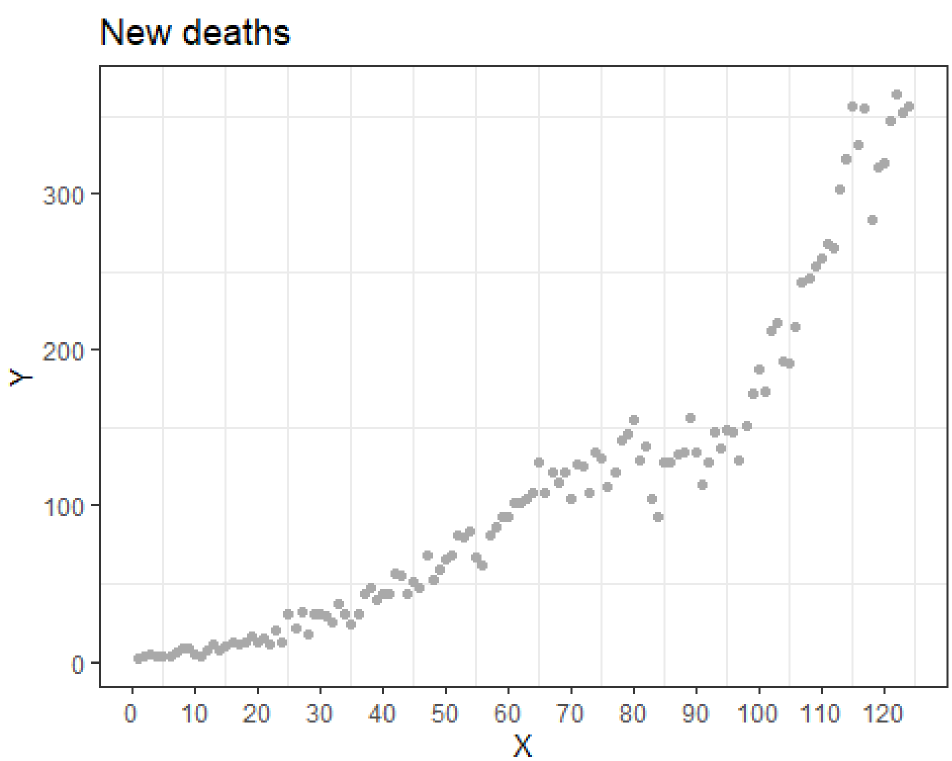
13/12/2021

Disposem de dades relacionades amb la COVID-19. En particular, per a aquesta activitat s'utilitzarem les dades corresponents a Espanya, extretes de l'Organització Mundial de la Salut: WHO-COVID-19-global-data-SPAIN.csv, amb dades d'evolució de la COVID-19 des del 03/01/2020 al 03/09/2021. Els dies es numeraran de manera ascendent i consecutiva, sent el dia 1 el corresponent a la data d'inici.

1 Modelització de la mortalitat

El creixement observat en les corbes de nous contagis i morts de la COVID-19 és clarament no lineal, creixent molt més ràpid a mesura que passen els dies. Es demana:

1. Utilitzar alguna de les tècniques de linealització per a ajustar les dades de noves morts (etiqueta New deaths) observats entre les dates 15/07/2020 i 15/11/2020, corresponents a la segona ona de pandemia. Justificar l'elecció.



Segons el model SIR, publicat per Kermack i McKendrick al 1927, i que pretèn explicar la dinàmica d´una malaltia infecciosa que s´estèn a una població susceptible, la taxa de mortalitat dels infectats és constant. Això fa esperar que les morts tinguin un creixement similar a les infeccions (noves, no acumulades).

Si cada individu infectat provoca un nombre d´infeccions secundàries (*basic reproductive ratio*, Ro), podríem suposar (si no s´apliquen mesures externes de control) que el creixement de les infeccions (i, paral.lelament, de les morts) seria no-lineal, potser exponencial.

Però tampoc podem descartar l´equació potència, que es fa servir precisament quan no coneixem la fòrmula exacta.

L´equació de la taxa de creixement de saturació podria ser addient si parlèssim del casos acumulats, i la pandèmia estigués molt avançada o no intervinguessin altres factors com, per exemple, que s´apliquessin mesures externes. La saturació de la corba de manera natural, voldria dir que la població susceptible hauria estat totalment infectada (els prèviament infectats es consideren inmunes al model SIR). Si mirem la gràfica de la corba, sembla que no s´hagi arribat encara al punt de saturació.

Mirarem quin model s´ajusta millor gràficament:

Model exponencial

$$y = \alpha e^{\beta x}$$

Si transformem l´expressió amb logaritmes:

log(Y) = log(a) + b*X

o el que és el mateix:

lny = a0 + a1*X

a0 = log(a)

a1 = b

a = exp(a0)

lny= log(Y)

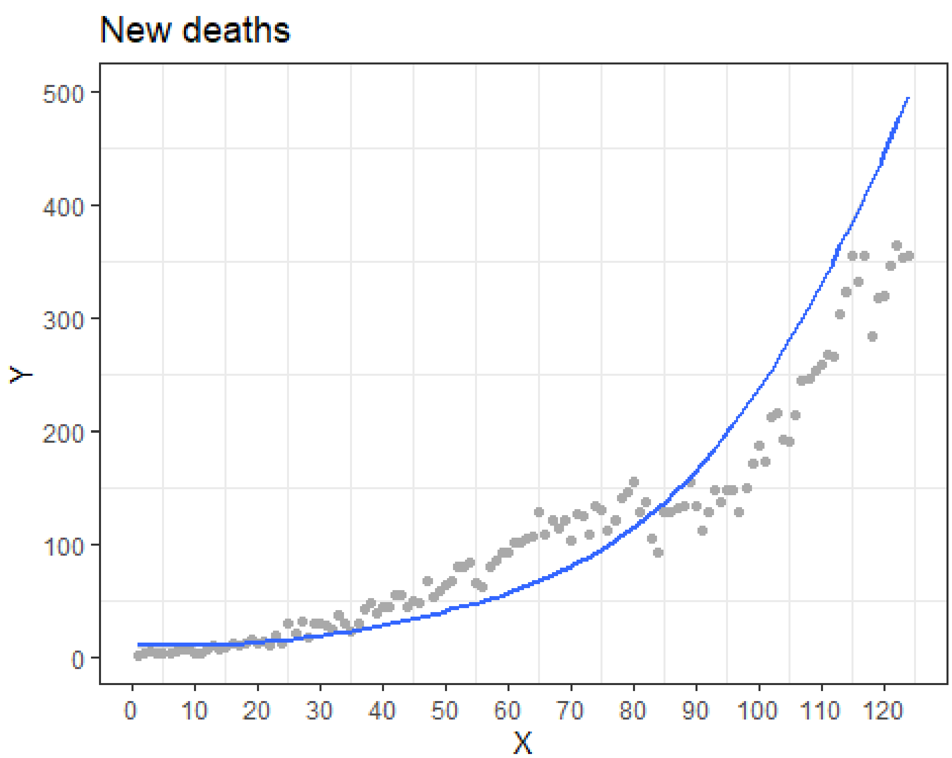
Amb les dades transformades, fem servir la regressió per mínims quadrats per trobar la corba que més s´ajusta:

```
##
## Formula: lny ~ log(a) + b * X
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a  7.46750    0.58066   12.86   <2e-16 ***
## b  0.03420    0.00108   31.67   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4303 on 122 degrees of freedom
##
## Number of iterations to convergence: 5
## Achieved convergence tolerance: 2.1e-07
```

Estimem la bondat d´ajust amb la correlació entre les prediccions del model i la variable resposta:

```
## [1] 0.9442319
```



Equació potència

$$y = \alpha x^{\beta}$$

De nou, transformem, ara amb log10:

log10(Y)= log10(a2) + b2*log10(X)

o el que és el mateix:

log10(Y)= a0 + a1*log10(X)

a0= log10(a2)

a1= b2

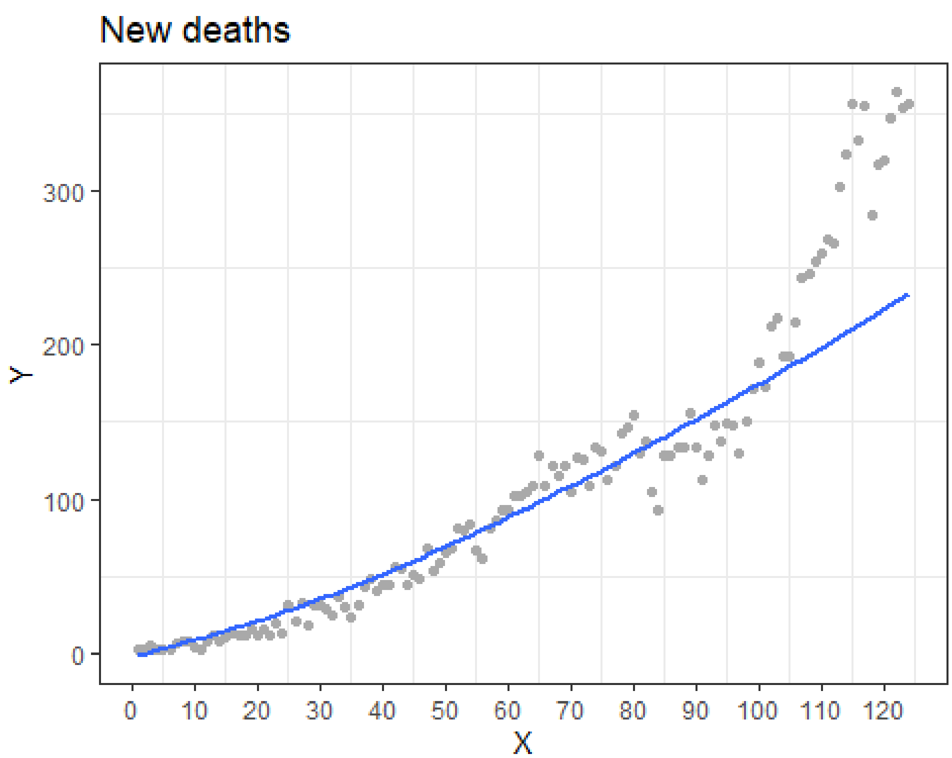
logy = log10(Y)

logx = log10(X)

```
##
## Formula: logy ~ log10(a2) + b2 * logx
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## a2  0.36979    0.04975   7.434 1.61e-11 ***
## b2  1.33672    0.03398  39.337  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1534 on 122 degrees of freedom
##
## Number of iterations to convergence: 4
## Achieved convergence tolerance: 4.179e-06
```

Estimem la bondat d´ajust:

```
## [1] 0.9627664
```



Equació de la taxa de creixement de saturació

$$y = \alpha \frac{x}{\beta + x}$$

Fem la transformació:

1/y = 1/a3 + (b3/a3)*(1/x)

O el que és el mateix:

1/y = a0 + a1*(1/X)

a0 = 1/a3

a1 = b3/a3

invy = 1/Y

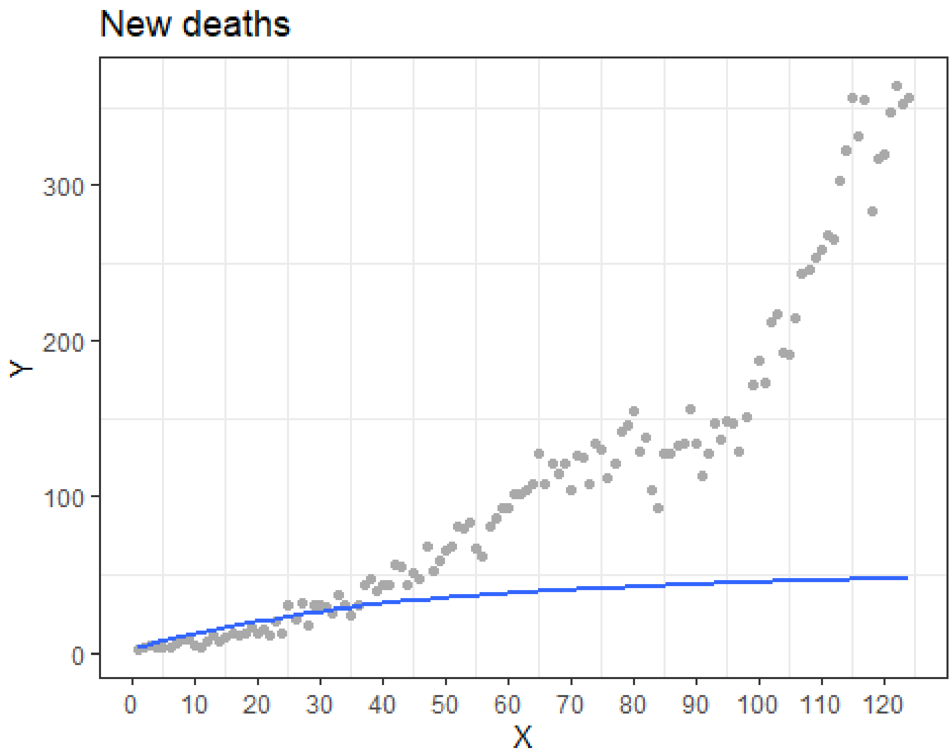
invx = 1/X

```
##
## Formula: invy ~ 1/a3 + (b3/a3) * invx
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## a3    63.88      18.87   3.384  0.00096 ***
## b3    41.45      13.43   3.085  0.00252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04766 on 122 degrees of freedom
##
## Number of iterations to convergence: 10
## Achieved convergence tolerance: 5.097e-08
```

Estimem la bondat d´ajust:

```
## [1] 0.8248809
```

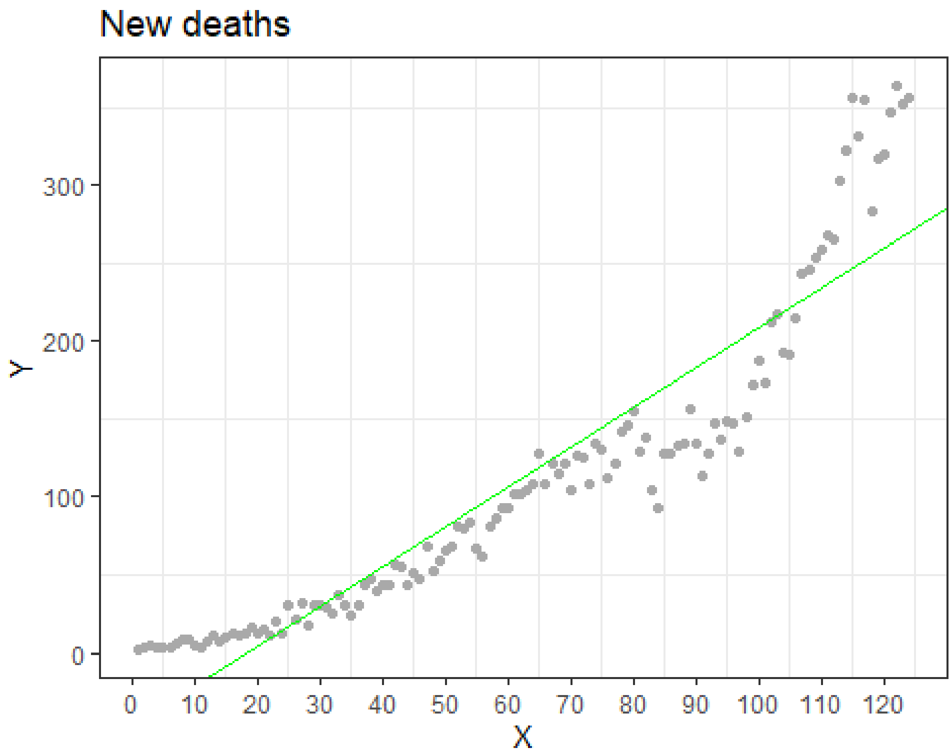
No és tan bona com les anteriors.



Gràficament comprobem que differeix molt de la recta original. Descartem aquesta opció. Qualsevol dels altres dos mètodes és acceptable.

2. Realitzar el mateix ajust mitjançant regressió lineal (n = 1).

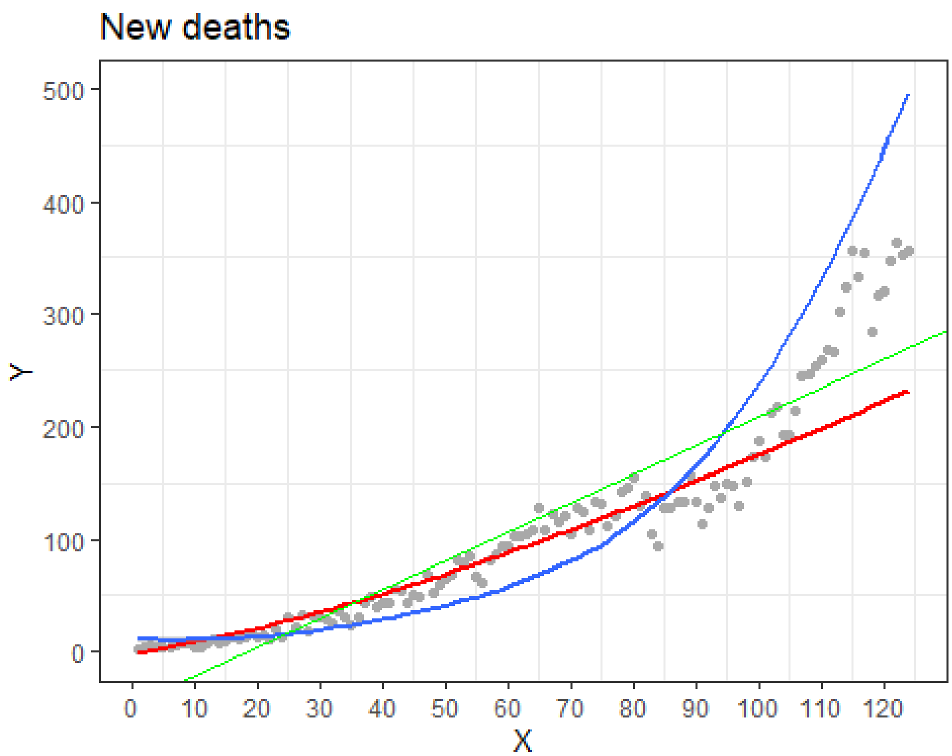
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.657 -20.775  -7.137  18.588 109.012
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -47.30488    6.64916  -7.114 8.36e-11 ***
## X             2.55907    0.09232  27.720  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.8 on 122 degrees of freedom
## Multiple R-squared:  0.863,  Adjusted R-squared:  0.8619
## F-statistic: 768.4 on 1 and 122 DF,  p-value: < 2.2e-16
```



Estimem la bondat d´ajust:

```
## [1] 0.9289691
```

3. Representar gràficamente sobre les dades originals els ajustos obtinguts mitjançant els models anteriors. Comentar raonadament els resultats obtinguts.



Fins aproximadament X=90, sembla que la transformació a potència (vermell) s´ajusta molt bé a les dades. També la regressió lineal(verd), fins als 80 dies.

Després sembla que hi ha un canvi, el creixement augmenta molt més ràpid, i el model exponencial sembla el més addient.

4. Determinar el millor model en base al càlcul del coeficient de correlacio r. Comentar raonadament els resultats obtinguts.

```
## Coeficient correlació ajust lineal: 0.9289691

## Coeficient correlació potència: 0.9099233

## Coeficient correlació exponencial: 0.8862745
```

El **coeficient de correlació**, r, és l´arrel quadrada del **coeficient de determninació**. Si r= 1 (Sr =0), tenim que la línia recta de la regressió lineal o la transformació, en cada cas, explica perfectament el 100% de la variabilitat de les dades. Si r=0 (Sr = St), l´aproximació no millora la mitjana a l'hora de descriure les dades.

El coeficient de correlació és més gran quan es fa l´ajust lineal. Al primer tram sí que s´ajusta molt bé a les dades, però hi ha un canvi de tendència després.

Els resultats dependran del tram analitzat de la corba: probablement, si augmentéssim la mida de X, el model lineal no seria el que millor per explicar el comportament de les dades.

5. Un dels usos més habituals dels models estadístics és el de predir possibles escenaris futurs. Emprant els models ajustats en els apartats anteriors, predir el numero de morts que s'aconseguirien 10 dies després de l'última data utilitzada en cas de deixar l'evolucio de la pandèmia sense control, és a dir, en cas de no aplicar cap mesura que mitigui la seva propagacio. Comentar raonadament els resultats obtinguts.

```
## Predicció amb funció potència 257.8105

## Predicció amb exponencial 729.7059

## Predicció amb regressió lineal 295.6107
```

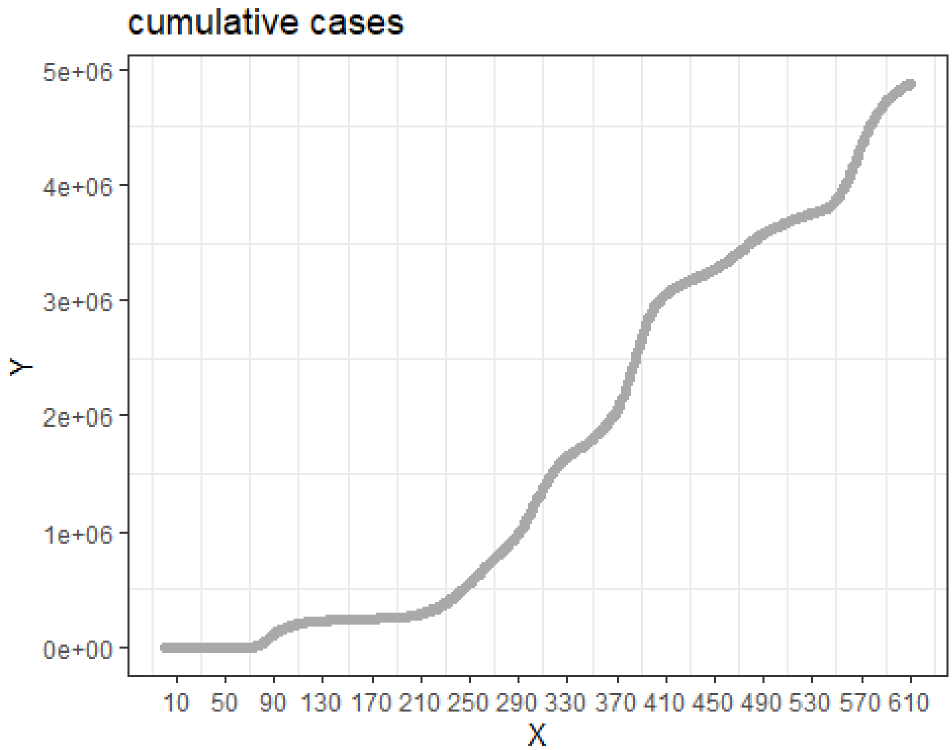
Amb el model exponencial el creixement és molt gran, i la predicció és molt més alta. Amb el model lineal, el creixement és lineal, per tant, més lent.

Donada la complexitat de les corbes de contagis (o morts en aquest cas), determinada per moltes variables, probablement el creixement no és purament exponencial, i el millor model seria la regressió lineal múltiple (que no veurem aquí). Com hem dit abans, no només s´ha de tenir en compte Ro, el model SIR determina que la transmissió de la infecció està controlada per un terme bilinear (depèn de la infectivitat del virus i de la susceptibilitat de la població).

2 Detectar els canvis de tendència

La regressió segmentada (o per segments) és una tècnica estadística que consisteix a separar les dades disponibles atesa l'observació de relacions lineals en diferents trams de dades. Aquesta tècnica en molt útil per a detectar els punts en els quals es produeix un salt brusc en la magnitud observada o un canvi de tendència en l'evolució de les dades.

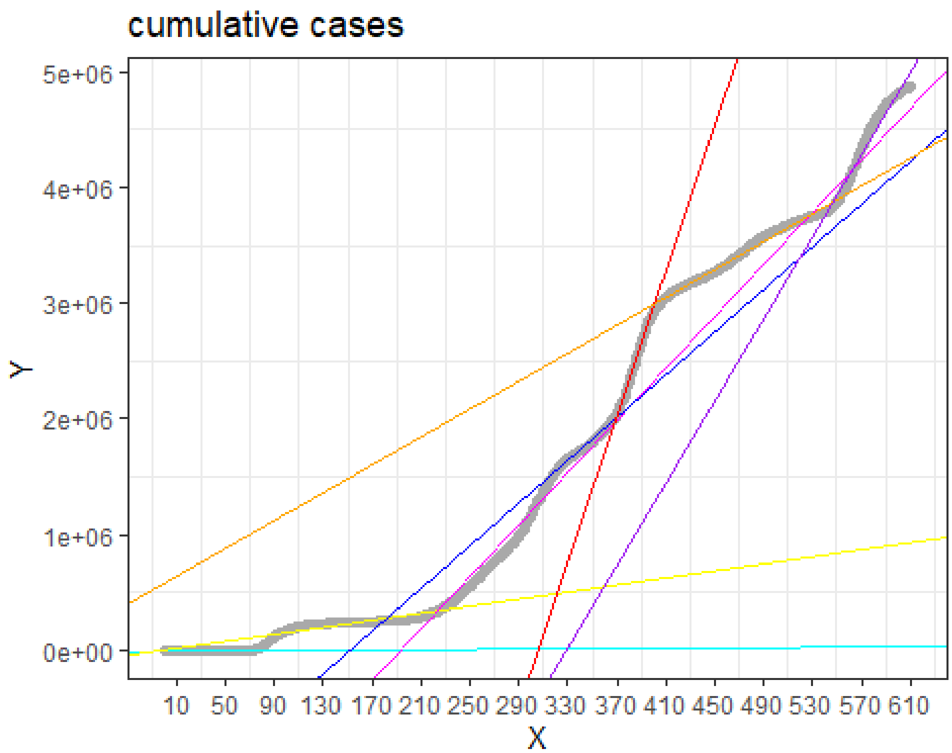
1. Emprar la regressio segmentada sobre les dades (tots els disponibles, és a dir, des del 03/01/2020 fins al 03/09/2021) de la corba de contagis acumulats (etiqueta Cumulative cases). Seleccionar un nombre de segments i els rangs de dades que s'assignen a cadascun simplement mitjançant observació.



Fem 7 segmentacions de la corba:

```
ini_segmentos = c(1,75,200,330,370,400,550)
fin_segmentos = c(75,200,330,370,400,550,610)
```

2. Representar gràficament la regressio segmentada obtinguda sobre les dades. Comentar raonadament els resultats obtinguts.

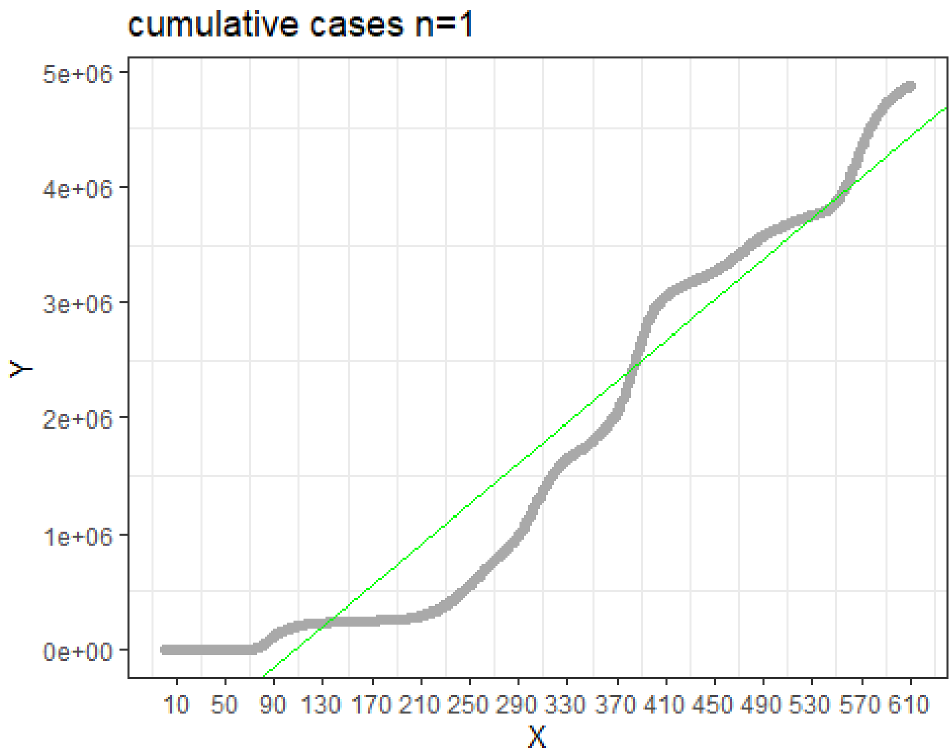


S´observa que hi ha moltes interseccions: hi ha molts canvis de pendent a la corba.

3. Realitzar l'ajust de la mateixa corba emprant regressió lineal bàsica (n = 1) i regressió lineal polinòmica de grau 2 (n = 2). Representar aquests dos ajustos sobre les dades i comparar els resultats obtinguts amb la regressió segmentada.

Ajust per n=1:

Representem gràficament:



Ara per n=2:

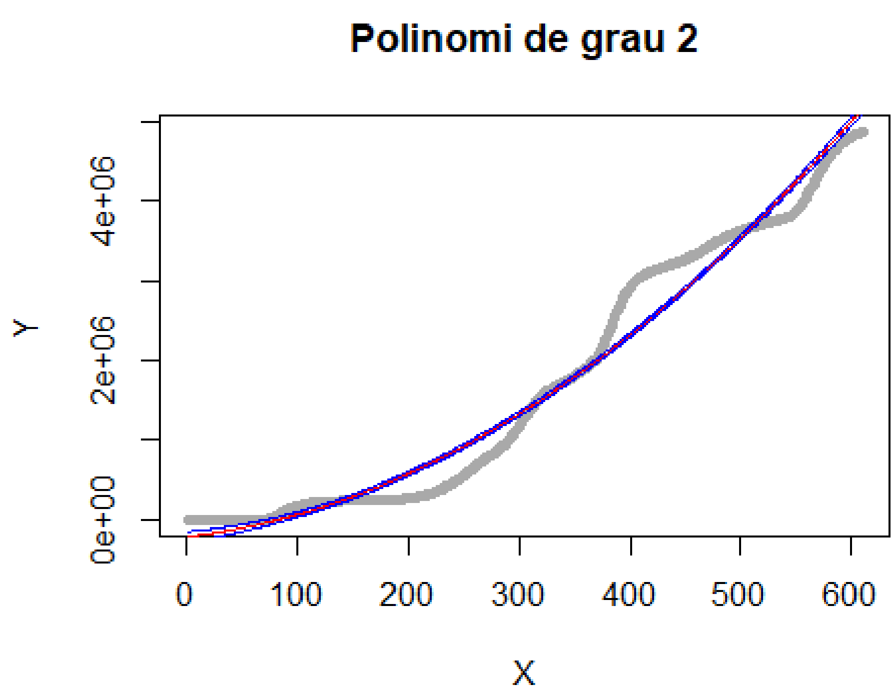
Si comparem els dos models, veiem que l´error residual stàndard és més petit en el model de grau 2. A més, R^2 és més gran: el model de grau 2 explica el 97,31% de la variació de la variable resposta, tot i que al model de grau també és alt (>93%):

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -706948 -331746   86303  298060  942194
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -951019.76   34293.70  -27.73   <2e-16 ***
## X              8825.66      97.25   90.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 423000 on 608 degrees of freedom
## Multiple R-squared:  0.9312, Adjusted R-squared:  0.9311
## F-statistic:  8235 on 1 and 608 DF,  p-value: < 2.2e-16

##
## Call:
```

```
## lm(formula = Y ~ poly(X, degree = 2, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397233 -225987   16691  137002  629422
##
## Coefficients:
##                                    Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                      -2.103e+05  3.224e+04  -6.521 1.47e-
10 ***
## poly(X, degree = 2, raw = TRUE)1  1.563e+03  2.437e+02   6.415 2.84e-
10 ***
## poly(X, degree = 2, raw = TRUE)2  1.189e+01  3.862e-01  30.774  < 2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 264600 on 607 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9731
## F-statistic: 1.1e+04 on 2 and 607 DF,  p-value: < 2.2e-16
```

els p-value obtinguts per a l'estadístic F són molt baixos(significatius), cosa que indica que els predictors introduïts al model estan relacionats amb la variable resposta.



Com més gran sigui el grau del polinomi més flexibilitat tindrà el model però, alhora, més risc d'overfitting. D'acord amb el principi de parsimònia, el grau òptim és el grau més baix que permeti explicar la relació entre les dues variables.

Per identificar-lo, es pot recórrer al contrast d'hipòtesis per anova:

```
## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ poly(X, degree = 2, raw = TRUE)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    608 1.0878e+14
## 2    607 4.2488e+13  1 6.6289e+13 947.03 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-value de la comparació entre el model lineal (model 1) i el quadràtic (model 2) és pràcticament zero (< 2.2e-16), indicant que el model lineal no és suficient.

La regressió segmentada s´ajusta més als punts de la corba, i pot ajudar quan no és fàcil trobar un model que expliqui les dades. El problema és que es pot estar perdent la naturalesa de la relació, és a dir, el model pot perdre explicació de la variabilitat. De la mateixa manera, al llibre Introduction to Statistical Learning desaconsellen l'ús de models polinòmics amb grau major de 3 o 4 a causa d'un excés de flexibilitat (overfitting), principalment als extrems del predictor X.

4. A partir d'aquest estudi, determinar els instants en els quals es produeix un canvi de tendència en l'evolució dels contagis, és a dir, els punts de tall de les rectes obtingudes per a dos segments consecutius. Comentar raonadament els resultats obtinguts.

La intersecció entre 2 rectes consecutives es dóna quan:

a1+b1x=a2+b2x

Intersecció recta 1 i 2:

```
## (Intercept)
##  -0.2416808
```

i les altres:

```
## (Intercept)
##    224.1607

## (Intercept)
##    383.1676

## (Intercept)
##    369.5236
```

```
## (Intercept)
##    401.3065

## (Intercept)
##    545.4739
```

També es pot fer amb el paquet segmented de R:

```
##
##  ***Regression Model with Segmented Relationship(s)***
##
## Call:
## segmented.lm(obj = lm.n1, seg.Z = ~X, psi = NA)
##
## Estimated Break-Point(s):
##             Est. St.Err
## psi1.X   72.822  2.162
## psi2.X  107.307  2.277
## psi3.X  204.648  4.552
## psi4.X  228.741  3.540
## psi5.X  256.483  3.454
## psi6.X  374.666  0.748
## psi7.X  401.387  0.589
## psi8.X  459.279  6.080
## psi9.X  493.057  3.621
## psi10.X 543.350  0.870
##
## Meaningful coefficients of the linear terms:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -537.33    7820.44  -0.069    0.945
## X               22.93     186.19   0.123    0.902
## U1.X          6215.92     580.22  10.713       NA
## U2.X         -5658.62     562.28 -10.064       NA
## U3.X          3267.85     975.52   3.350       NA
## U4.X          5175.00    1235.94   4.187       NA
## U5.X          3995.11     773.28   5.166       NA
## U6.X         18615.87     816.12  22.810       NA
## U7.X        -25917.87     851.18 -30.449       NA
## U8.X          2342.93     629.10   3.724       NA
## U9.X         -4008.29     658.01  -6.092       NA
## U10.X        14093.59     382.83  36.814       NA
##
## Residual standard error: 32830 on 588 degrees of freedom
## Multiple R-Squared: 0.9996,  Adjusted R-squared: 0.9996
##
## Convergence attained in 21 iter. (rel. change 2.9055e-06)
```

Punts de tall:

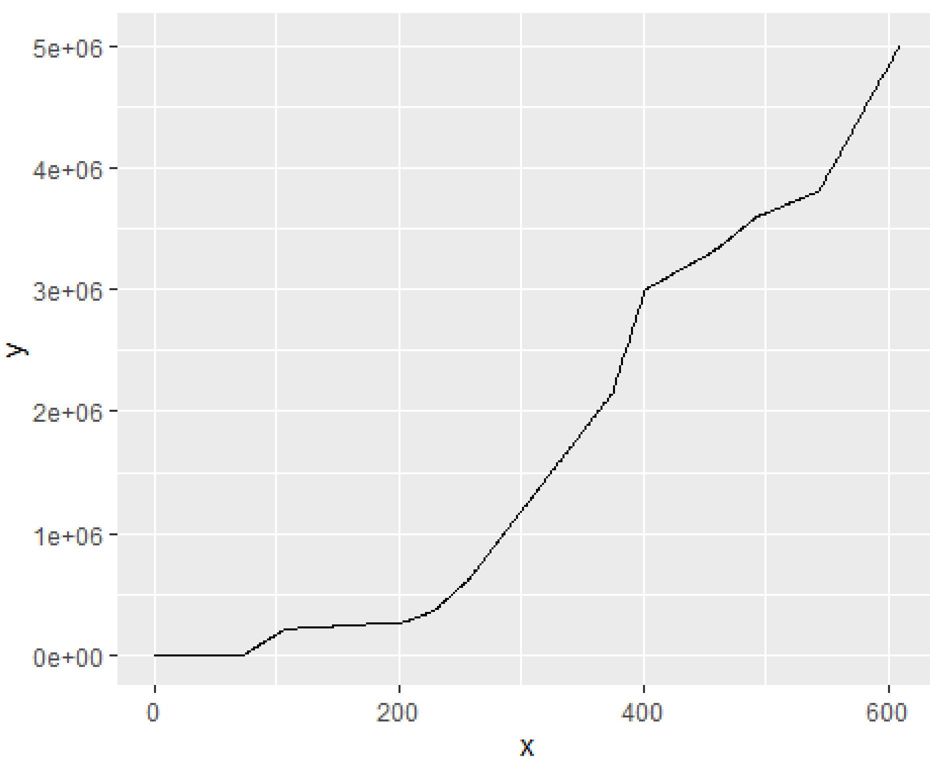
```
##           Initial      Est.    St.Err
## psi1.X   56.36364  72.82153 2.1621553
```

```
## psi2.X  111.72727 107.30710 2.2767947
## psi3.X  167.09091 204.64780 4.5520318
## psi4.X  222.45455 228.74134 3.5400230
## psi5.X  277.81818 256.48268 3.4542019
## psi6.X  333.18182 374.66557 0.7475418
## psi7.X  388.54545 401.38726 0.5885570
## psi8.X  443.90909 459.27857 6.0801725
## psi9.X  499.27273 493.05742 3.6205863
## psi10.X 554.63636 543.34976 0.8698950
```

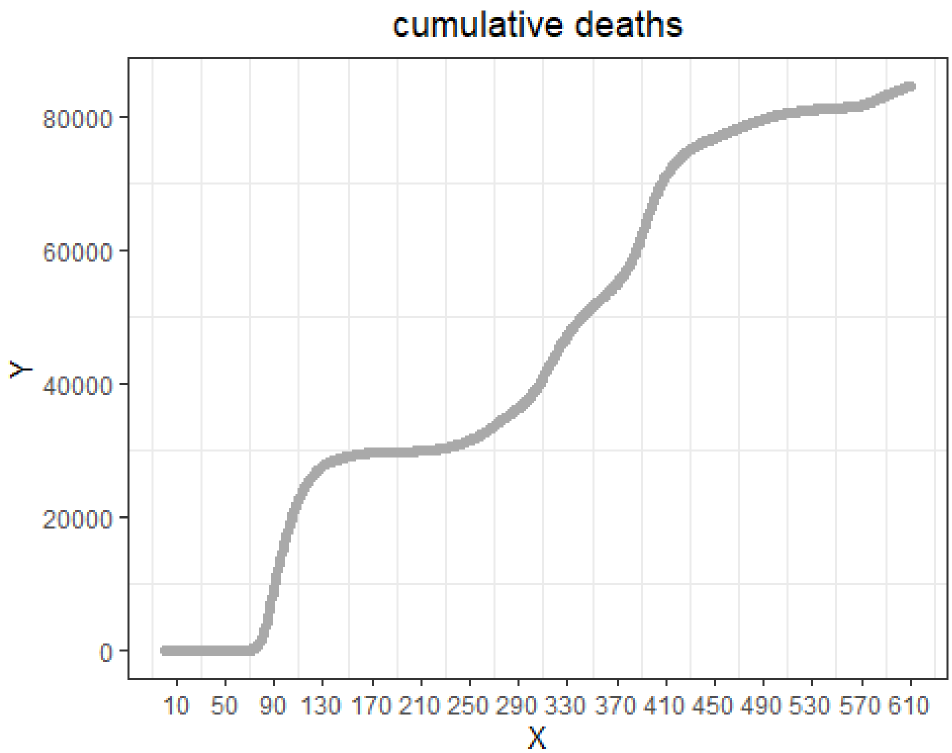
I les pendents:

```
## $X
##              Est. St.Err.   t value CI(95%).l CI(95%).u
## slope1     22.931 186.190   0.12316   -342.75    388.61
## slope2   6238.800 549.530  11.35300   5159.60   7318.10
## slope3    580.230 119.060   4.87320    346.38    814.07
## slope4   3848.100 968.230   3.97430   1946.50   5749.70
## slope5   9023.100 768.170  11.74600   7514.40  10532.00
## slope6  13018.000  88.738 146.70000  12844.00  13192.00
## slope7  31634.000 811.280  38.99300  30041.00  33227.00
## slope8   5716.200 257.540  22.19600   5210.40   6222.00
## slope9   8059.100 573.970  14.04100   6931.80   9186.40
## slope10  4050.800 321.770  12.58900   3418.90   4682.80
## slope11 18144.000 207.420  87.47600  17737.00  18552.00
```

Representació gràfica dels segments:



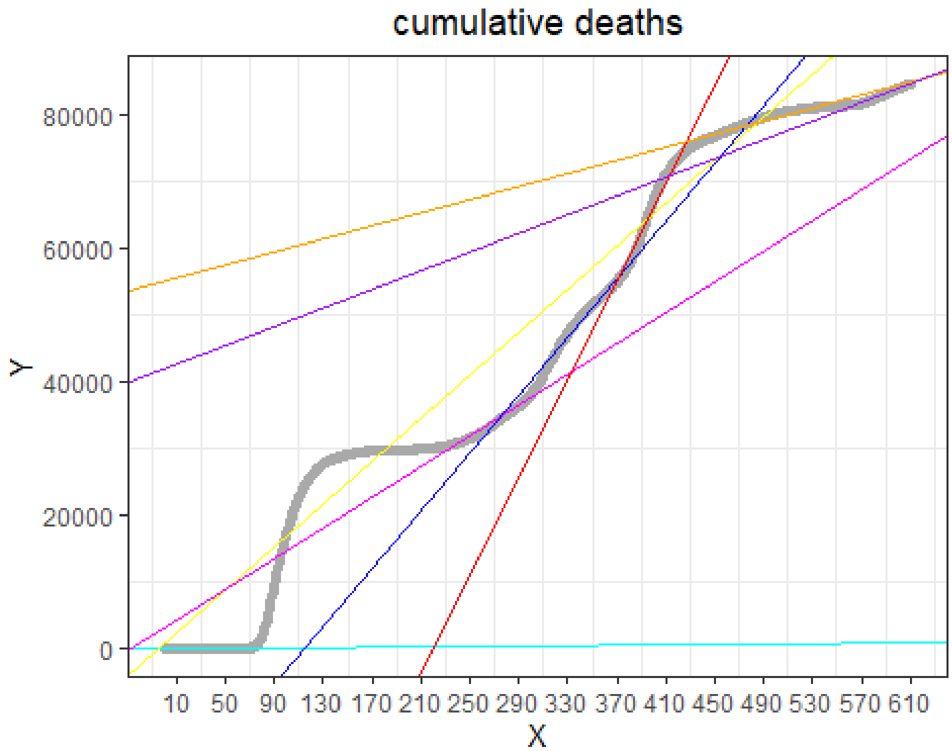
1B. Repetir els apartats anteriors emprant les dades disponibles de la corba de morts acumulades (etiqueta Cumulative deaths)



ini_segmentos = c(1,75,220,310,380,430,560)

fin_segmentos = c(75,220,310,380,430,560,610)

2B. Regressió segmentada de “Cumulative_deaths”:

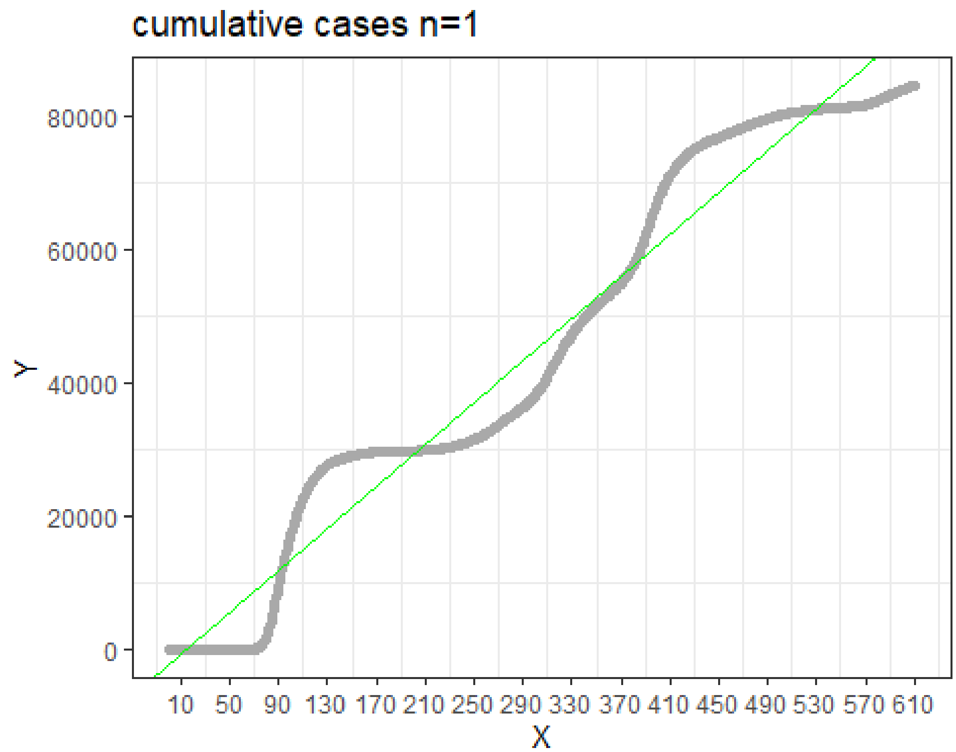


Igual que a l´apartat anterior, s´observen molts canvis de pendent.

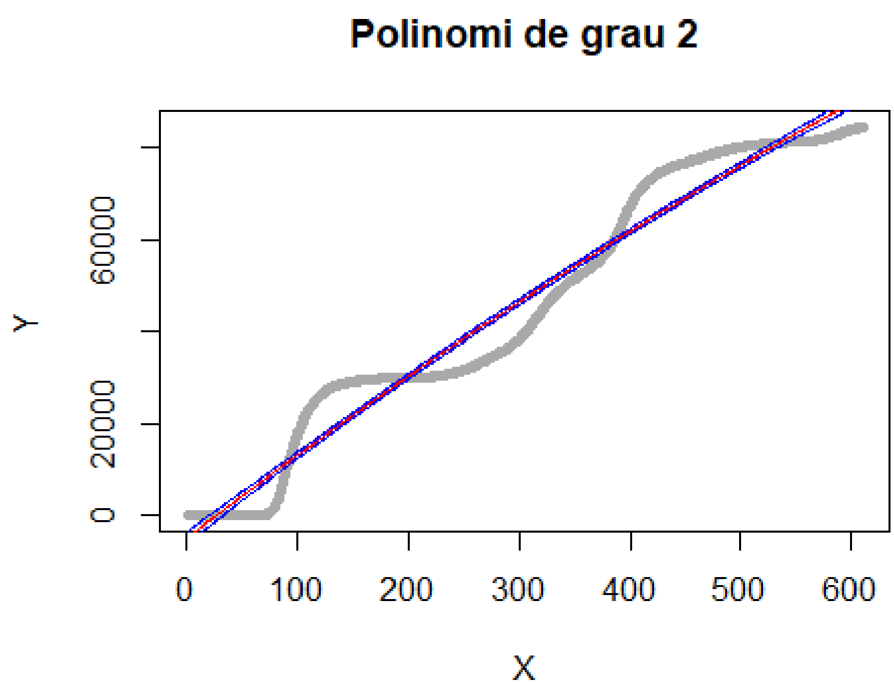
3B. Realitzar l’ajust de la mateixa corba emprant regressió lineal bàsica (n = 1) i regressió lineal polinòmica de grau 2 (n = 2). Representar aquests dos ajustos sobre les dades i comparar els resultats obtinguts amb la regressió segmentada.

Ajust per n=1:

Representem gràficament:


cumulative cases n=1
80000
60000
40000
20000
0
Y
10 50 90 130 170 210 250 290 330 370 410 450 490 530 570 610
X

Ara per n=2:


Polinomi de grau 2
0
20000
60000
Y
0 100 200 300 400 500 600
X

Si comparem els dos models, veiem que l´error residual stàndard és una micamés petit en el model de grau 2, i R^2 és una mica més gran, però sense grans diferències:

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9377  -5205  -1062   5259   9608
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2276.276    465.273  -4.892 1.28e-06 ***
## X             157.857      1.319 119.635  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5739 on 608 degrees of freedom
## Multiple R-squared:  0.9593, Adjusted R-squared:  0.9592
## F-statistic: 1.431e+04 on 1 and 608 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = Y ~ poly(X, degree = 2, raw = TRUE))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8512.1 -4851.2  -905.7  5109.3  9535.0
##
## Coefficients:
##                                    Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                      -5.090e+03  6.830e+02  -7.453 3.15e-
13 ***
## poly(X, degree = 2, raw = TRUE)1  1.854e+02  5.162e+00  35.922  < 2e-
16 ***
## poly(X, degree = 2, raw = TRUE)2 -4.516e-02  8.182e-03  -5.519 5.05e-
08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5604 on 607 degrees of freedom
## Multiple R-squared:  0.9612, Adjusted R-squared:  0.9611
## F-statistic:  7518 on 2 and 607 DF,  p-value: < 2.2e-16
```

Els p-value obtinguts per a l'estadístic F són significatius, cosa que indica que els predictors introduïts al model estan relacionats amb la variable resposta.

Contrast d'hipòtesis per anova:

```
## Analysis of Variance Table
##
## Model 1: Y ~ X
## Model 2: Y ~ poly(X, degree = 2, raw = TRUE)
##   Res.Df        RSS Df Sum of Sq     F    Pr(>F)
## 1    608 2.0023e+10
## 2    607 1.9066e+10  1 956762301 30.46 5.053e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-value de la comparació entre el model lineal (model 1) i el quadràtic (model 2) és molt petit (5.053e-08), indicant que el model lineal no és suficient.

4. A partir d'aquest estudi, determinar els instants en els quals es produeix un canvi de tendència en l'evolució dels contagis, és a dir, els punts de tall de les rectes obtingudes per a dos segments consecutius. Comentar raonadament els resultats obtinguts.

La intersecció entre 2 rectes consecutives es dóna quan:

a1+b1x=a2+b2x

Intersecció recta 1 i 2:

```
## (Intercept)
##   -4.415254
```

i la resta:

```
## (Intercept)
##    57.28685

## (Intercept)
##    275.7779

## (Intercept)
##    372.2455

## (Intercept)
##    426.4522

## (Intercept)
##    625.8167
```

A R existeixen altres mètodes per fer regressió segmentada.

Paquet segmented de R:

```
##
##  ***Regression Model with Segmented Relationship(s)***
##
## Call:
```

```
## segmented.lm(obj = lm.n1, seg.Z = ~X, psi = NA)
##
## Estimated Break-Point(s):
##             Est. St.Err
## psi1.X   77.790  0.214
## psi2.X  106.769  0.388
## psi3.X  133.379  0.714
## psi4.X  234.211  3.709
## psi5.X  254.474  3.770
## psi6.X  285.886  1.384
## psi7.X  384.085  0.690
## psi8.X  406.561  0.837
## psi9.X  426.807  1.377
## psi10.X 483.234  2.686
##
## Meaningful coefficients of the linear terms:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -72.129     86.688  -0.832    0.406
## X              2.866      1.931   1.484    0.138
## U1.X         760.407      8.580  88.629       NA
## U2.X        -518.664     12.510 -41.462       NA
## U3.X        -227.642      9.395 -24.231       NA
## U4.X          50.621     14.662   3.452       NA
## U5.X          57.304     16.448   3.484       NA
## U6.X         113.485      7.678  14.780       NA
## U7.X         263.089     12.726  20.673       NA
## U8.X        -278.033     19.327 -14.386       NA
## U9.X        -144.397     14.917  -9.680       NA
## U10.X        -44.552      3.166 -14.070       NA
##
## Residual standard error: 376.6 on 588 degrees of freedom
## Multiple R-Squared: 0.9998,  Adjusted R-squared: 0.9998
##
## Convergence attained in 4 iter. (rel. change 1.0361e-14)
```

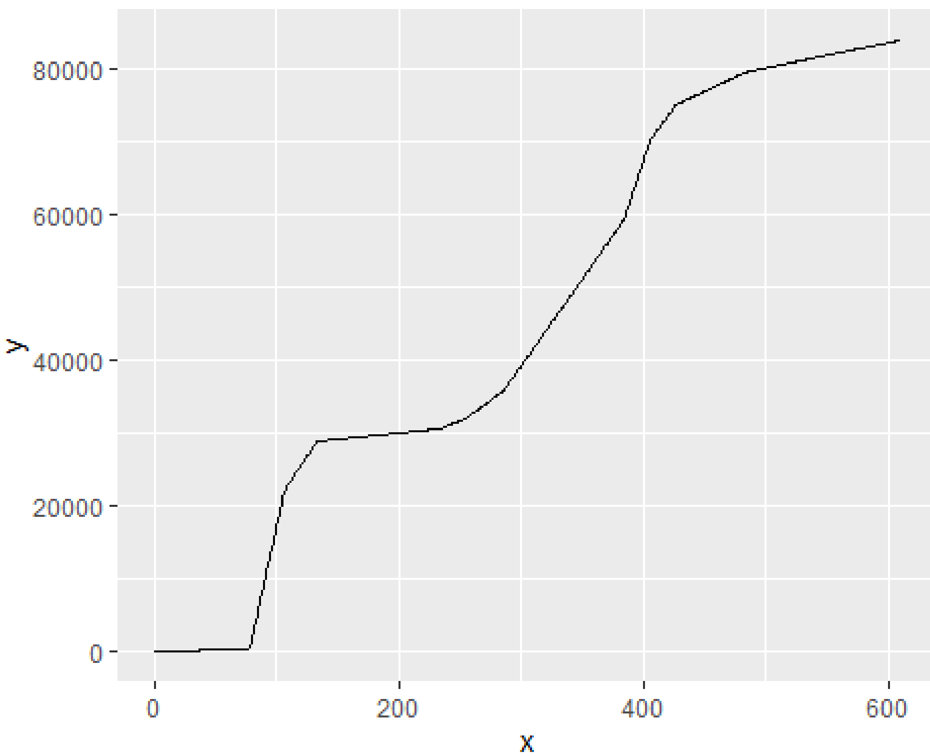
Punts de tall:

```
##           Initial     Est.    St.Err
## psi1.X   56.36364  77.7896 0.2139073
## psi2.X  111.72727 106.7694 0.3882474
## psi3.X  167.09091 133.3789 0.7136873
## psi4.X  222.45455 234.2108 3.7088376
## psi5.X  277.81818 254.4737 3.7703248
## psi6.X  333.18182 285.8859 1.3842246
## psi7.X  388.54545 384.0849 0.6903162
## psi8.X  443.90909 406.5606 0.8374559
## psi9.X  499.27273 426.8067 1.3772852
## psi10.X 554.63636 483.2336 2.6855271
```

I les pendents:

```
## $X
##             Est.  St.Err.  t value CI(95%).l CI(95%).u
## slope1    2.8664  1.93120   1.4843  -0.92639    6.6593
## slope2  763.2700  8.35950  91.3060 746.86000  779.6900
## slope3  244.6100  9.30620  26.2850 226.33000  262.8900
## slope4   16.9670  1.28550  13.1990  14.44300   19.4920
## slope5   67.5880 14.60600   4.6275  38.90200   96.2740
## slope6  124.8900  7.56320  16.5130 110.04000  139.7500
## slope7  238.3800  1.32460 179.9600 235.78000  240.9800
## slope8  501.4700 12.65700  39.6190 476.61000  526.3300
## slope9  223.4300 14.60600  15.2980 194.75000  252.1200
## slope10  79.0360  3.03230  26.0650  73.08100   84.9920
## slope11  34.4850  0.91165  37.8260  32.69400   36.2750
```

Representació gràfica dels segments:



Mètode step functions:

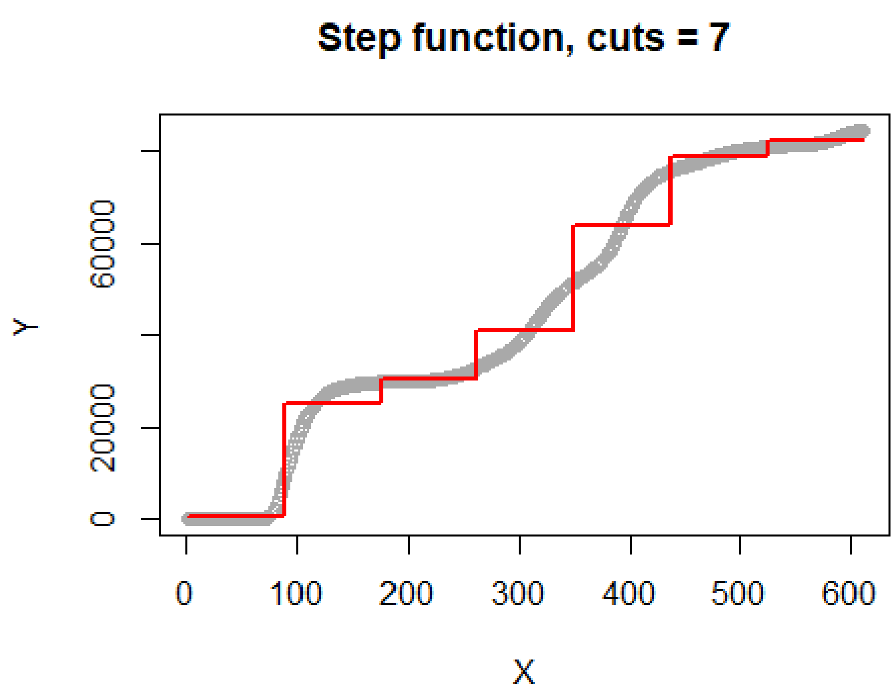
L'estratègia del mètode step functions consisteix a dividir el rang del predictor X en diversos subintervals i ajustar una constant diferent per a cadascú.

la funció cut() estableix n punts de tall i torna una variable qualitativa que indica el subinterval a què pertany cada observació:

```
##
## Call:
## lm(formula = Y ~ cut(X, 7))
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -16552.0   -991.0   -464.6   1887.5  11950.0 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           529.0      488.6   1.083    0.279    
## cut(X, 7)(88,175]   24628.1      693.0  35.537   <2e-16 ***
## cut(X, 7)(175,262]  29950.4      693.0  43.217   <2e-16 ***
## cut(X, 7)(262,349]  40551.4      693.0  58.513   <2e-16 ***
## cut(X, 7)(349,436]  63271.0      693.0  91.296   <2e-16 ***
## cut(X, 7)(436,523]  78341.9      693.0 113.042   <2e-16 ***
## cut(X, 7)(523,611]  81720.5      693.0 117.918   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4584 on 603 degrees of freedom
## Multiple R-squared:  0.9742, Adjusted R-squared:  0.974 
## F-statistic:  3797 on 6 and 603 DF,  p-value: < 2.2e-16
```

L'intercept (529) s'interpreta com els casos que hi ha de mitjana per sota del dia 88, i el coeficient de regressió estimat de cada grup, com l´increment mitjà de casos.



Ajust local del model polinòmic: loess()

El paràmetre a ajustar en regressió local és l'span (% d'observacions veïnes que cal considerar a cada ajustament). Com més gran sigui l'span, més suau serà l'ajust.

El polinomi local emprat per ajustar cada subset de dades sol ser sempre de primer o segon grau, és a dir, o bé un ajustament lineal o bé un quadràtic. Encara que des del punt de vista teòric es poden emprar polinomis de major grau, aquests tendeixen a produir overfit i redueixen la precisió del model.