PEC 8 - Aproximación de funciones y regresión (II)

Fecha de entrega: 22/05/2022

Descripción del problema

Las epidemias y pandemias son objetos de estudio muy importantes ya que afectan directamente al desarrollo de las actividades habituales de una sociedad. Su tratamiento se basa en analizar las características particulares de una determinada enfermedad o infección, estudiar su propagación y proponer medidas para su control. Es evidente que los factores que influyen en la evolución de una pandemia están fuertemente relacionados entre ellos, contando a su vez con un gran componente de incertidumbre. Esto implica que su tratamiento matemático sea altamente complejo, necesitando de una amplia variedad de herramientas numéricas y de una estudio multidisciplinar.

Las agencias de salud pública, tanto nacionales como internacionales, suelen recabar datos de muy diversa índole relacionados con la evolución de las epidemias (pandemias) que ocurren en todo el planeta. Por lo tanto, una gran cantidad de información heterogénea está disponible públicamente para su tratamiento. Basada en ella se pueden desarrollar modelos estadísticos que expliquen el comportamiento de epidemias pasadas, ayuden a desarrollar políticas de control para epidemias activas y permitan predecir posibles escenarios futuros.

Disponéis de datos relacionados con la pandemia de la COVID-19, una enfermedad producida por el virus SARS-CoV-2. En particular, para esta actividad se utilizan los datos correspondientes a España, extraídos de la Organización Mundial de la Salud. Se os proporciona el fichero *WHO-COVID-19-global-data-SPAIN.csv*, con datos de evolución de la COVID-19 desde el 03/01/2020 al 03/09/2021. Para importar los datos se os proporciona un *script* de R, que facilita la correcta lectura del fichero anterior, dada una fecha de inicio, una fecha de fin y una etiqueta. Los días se numerarán de manera ascendente y consecutiva, siendo el día 1 el correspondiente a la fecha de inicio.

1. Modelización de la mortalidad

El crecimiento observado en las curvas de nuevos contagios y muertes de la COVID-19 es claramente no lineal, creciendo mucho más rápido a medida que pasan los días. Se pide:

1. Utilizar alguna de las técnicas de linealización para ajustar los datos de nuevas muertes (etiqueta *New_deaths*) observados entre las fechas 15/12/2020 y 01/02/2021, correspondientes a la tercera ola de pandemia en España. Justificar la elección.
2. Realizar el mismo ajuste mediante regresión lineal ($n = 1$).
3. Representar gráficamente sobre los datos originales los ajustes obtenidos mediante los modelos

anteriores. Comentar razonadamente los resultados obtenidos.

4. Determinar el mejor modelo en base al cálculo del coeficiente de correlación r. Comentar razonadamente los resultados obtenidos.

5. Uno de los usos más habituales del los modelos estadísticos es el de predecir posibles escenarios futuros. Empleando los modelos ajustados en los apartados anteriores, predecir el numero de muertes que se alcanzarían 15 días después de la última fecha utilizada en caso de dejar la evolución de la pandemia sin control, es decir, en caso de no aplicar ninguna medida que mitigue su propagación. Comentar razonadamente los resultados obtenidos.

Solución Ejercicio 1

1. Observando los datos, parece que estos siguen un crecimiento exponencial. Por ello, es razonable asumir que un modelo exponencial, del tipo $y = \alpha \exp \beta x$, es adecuado para realizar el ajuste. Como se puede ver en la página 40 de la guía, este modelo se puede linealizar como $\log(y) = \log(\alpha) + \beta x$, lo que permite obtener los coeficientes mediante regresión lineal. Entonces, los coeficientes $c_0 = \log(\alpha)$ y $c_1 = \beta$ se pueden calcular resolviendo el sistema de ecuaciones normales o, directamente mediante las soluciones obtenidas en la página 37 de la guía.

   El sistema de ecuaciones normales que tenemos que resolver es:

$$\begin{aligned} 49c_0 + 1225c_1 &= 272.2909 \\ 1225c_0 + 40425c_1 &= 7112.379 \end{aligned}$$

   Solución:

$$c_0 = 4.77862360, \quad c_1 = 0.03113333.$$

   Por tanto, $\alpha = \exp(c_0) = 118.9405$ y $\beta = c_1 = 0.03113333$.

2. Realizamos el ajuste mediante regresión lineal, es decir, con un modelo del tipo $y = c_0 + c_1 x$. De nuevo, los coeficientes c_0 y c_1 se pueden calcular resolviendo el sistema de ecuaciones normales o, directamente mediante las soluciones obtenidas en la página 37 de la guía.

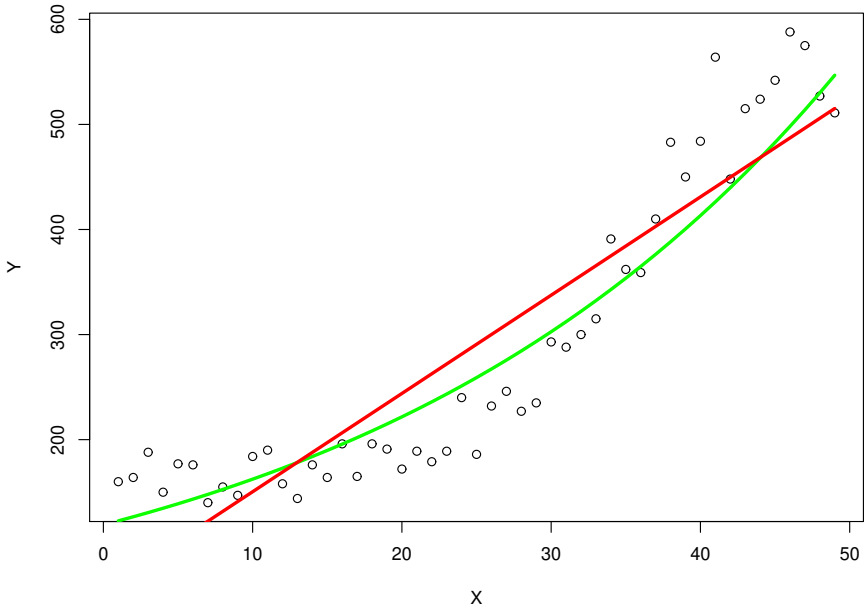
   El sistema de ecuaciones normales que tenemos que resolver es:

$$\begin{aligned} 49c_0 + 1225c_1 &= 14245 \\ 1225c_0 + 40425c_1 &= 447779 \end{aligned}$$

   Solución:

$$c_0 = 56.903061, \quad c_1 = 9.352449.$$

3. Para la representación gráfica se pueden utilizar los comandos de R, *plot* y/o *lines*. En la siguiente gráfica se representan los ajustes obtenidos en los puntos anteriores. De forma visual, parece que el modelo exponencial representa de una manera más fiable los datos.



Ajustes mediante un modelo exponencial (verde) y un modelo de regresión lineal (rojo).

4. Calculamos los coeficientes de correlación (expresión (31) de la guía) para cada uno de los ajustes anteriores:

$$r_E = \sqrt{\frac{S_t - S_{r,E}}{S_t}} = 0.9490057, \quad \text{donde} \quad S_{r,E} = \sum_{i=1}^{m} (y_i - \alpha \exp(\beta x_i))$$

$$r_L = \sqrt{\frac{S_t - S_{r,L}}{S_t}} = 0.9106591, \quad \text{donde} \quad S_{r,L} = \sum_{i=1}^{m} (y_i - c_0 - c_1 x_i)$$

Observamos que, en términos del coeficiente de correlación, el modelo exponencial se comporta mejor que el modelo lineal.

5. Para hacer una predicción, basta con evaluar los modelos ajustados en el tiempo que nos piden, en este caso, 15 días después de la última fecha empleada, que se corresponde con $x = 49$. Si denotamos la fecha de la predicción como $x_p = 64$, tendríamos que $y_p = \alpha \exp(\beta x_p) = 872.32$ para el modelo exponencial y $y_p = c_0 + c_1 x_p = 655.46$ para el modelo de regresión lineal. Claramente, parece que el modelo lineal infraestima de manera considerable.

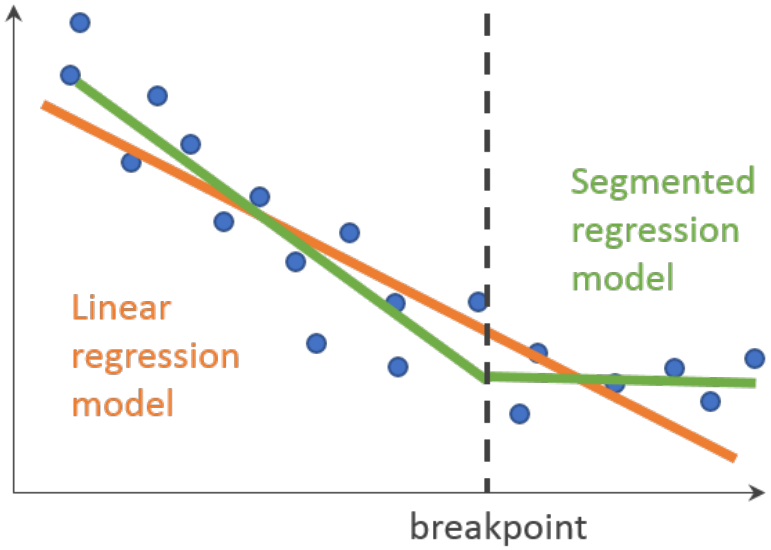
Código R

```
####### Funciones auxiliares #######
source('Lectura_datos_por_fecha.R')

# Contruccion de la matriz de los datos evaluados en las funciones base (Phi)
myPhi = function(x, n){

  Phi = matrix(1, length(x), n+1)
  for (i in 1:n) {
    Phi[, i+1] = x^i
  }
  return(Phi)

}

# Funcion que resuelve el sistema de ecuaciones normales mediante la expresion (5)
    de la guia
mylssolve = function(A, b){

  AT = t(A)
  return(solve(AT%*%A, AT%*%b))

}

# Funcion que realiza un ajuste lineal polinomico de grado "n". Devuelve los
    coeficientes del polinomio.
mypolyfit = function(x, y, n){

  Phi = myPhi(x, n)

  c = mylssolve(Phi, y)
  return(c)

}

# Funcion que evalua un ajuste polinomico definido por los coeficientes "c" en los
    puntos "x" (que puede ser un vector de puntos)
myeval = function(x, c){

  f = 0
```

```
  for (i in 1:length(c)) {
    f = f + c[i]*x^(i-1)
  }
  return(f)
}


######### Ejercicio 1 #########
## Lectura de los datos
Y = myReadData_byDate('WHO-COVID-19-global-data-SPAIN.csv', '15/12/2020', '01/02/
    2021', 'New_deaths')
m = length(Y)
X = 1:m

# Representacion de los datos
plot(X, Y)

## Modelo exponencial
logY = log(Y)
CkE = mypolyfit(X, logY, 1)
alpha = exp(CkE[1])
beta = CkE[2]

## Regresion lineal de grado 1
CkP = mypolyfit(X, Y, 1)

## Representacion grafica
lines(X, alpha*exp(beta*X), col='green', lwd=3)
lines(X, myeval(X, CkP), col='red', lwd=3)

## Calculo del coeficiente de determinacion para el modelo exponencial y polinomico
     de grado 1
St = sum( (Y - mean(Y))^2 )

Sr_E = sum( (Y - alpha*exp(beta*X))^2 )
Sr_P = sum( (Y - myeval(X, CkP))^2 )

r_E = sqrt( (St - Sr_E)/St )
r_P = sqrt( (St - Sr_P)/St )
print(r_E)
print(r_P)

## Prediccion
x15 = m+15
y15_E = alpha*exp(beta*x15)
y15_P = myeval(x15, CkP)
print(y15_E)
print(y15_P)
```

2. Detectar los cambios de tendencia

La regresión segmentada (o por segmentos) es una técnica estadística que consiste en separar los datos disponibles atendiendo a la observación de relaciones lineales en distintos tramos de datos. Esta técnica en muy útil para detectar los puntos en los que se produce un salto brusco en la magnitud observada o un cambio de tendencia en la evolución de los datos. En la siguiente figura podemos ver un ejemplo de regresión segmentada.



Se pide:

1. Emplear la regresión segmentada sobre los datos (todos los disponibles, es decir, desde el 03/01/2020 hasta el 03/09/2021) de la curva de contagios acumulados (etiqueta *Cumulative_cases*). Seleccionar un número de segmentos y los rangos de datos que se asignan a cada uno simplemente mediante observación.
2. Representar gráficamente la regresión segmentada obtenida sobre los datos. Comentar razonadamente los resultados obtenidos.
3. Realizar el ajuste de la misma curva empleando regresión lineal básica ($n = 1$) y regresión lineal polinómica de grado 2 ($n = 2$). Representar estos dos ajustes sobre los datos y comparar

los resultados obtenidos con la regresión segmentada.

4. A partir de este estudio, determinar los instantes en los que se produce un cambio de tendencia en la evolución de los contagios, es decir, los puntos de corte de las rectas obtenidas para dos segmentos consecutivos. Comentar razonadamente los resultados obtenidos.

5. Repetir los apartados anteriores empleando los datos disponibles de la curva de muertes acumuladas (etiqueta *Cumulative_deaths*).

Solución Ejercicio 2

1. En nuestro caso, hemos elegido 7 segmentos, cuyas fechas de inicio son son las que se corresponden con los días 1, 81, 111, 221, 371, 401 y 551, y sus fechas de fin se corresponden con los días 80, 110, 220, 370, 400, 550 y 610. Las aproximaciones para cada uno de los segmentos son:

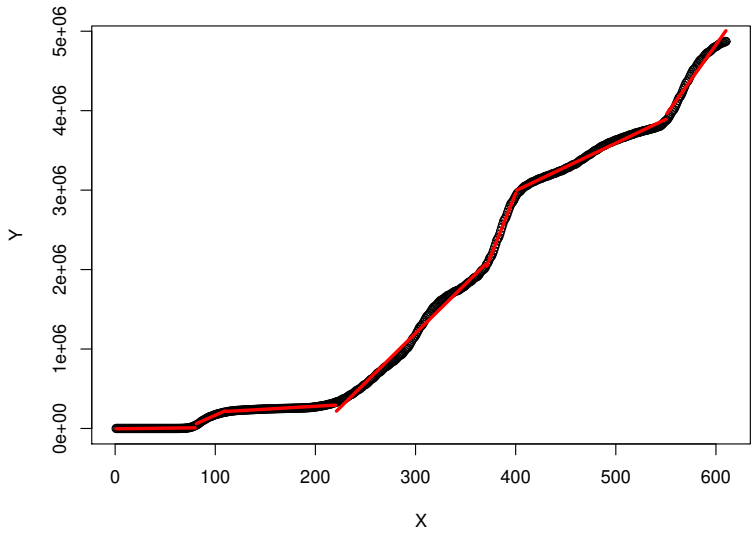
   - Segmento 1: $\hat{y} = -4028.41 + 156.48x$.
   - Segmento 2: $\hat{y} = -393730.31 + 5602.01x$.
   - Segmento 3: $\hat{y} = 132402.73 + 734.91x$.
   - Segmento 4: $\hat{y} = -2527197.78 + 12420.14x$.
   - Segmento 5: $\hat{y} = -9686157.66 + 31611.89x$.
   - Segmento 6: $\hat{y} = 595025.95 + 5992.78x$.
   - Segmento 7: $\hat{y} = -5789256.61 + 17698.37x$.

2. En la siguiente gráfica podemos se representan los segmentos obtenidos sobre los datos. Observamos que la regresión segmentada ajusta de forma óptima los datos.

3. Las aproximaciones obtenidas son:

   - Regresión lineal ($n = 1$): $\hat{y} = -951019.76 + 8825.66x$.
   - Regresión polinómica ($n = 2$): $\hat{y} = -210263.91 + 1563.34x + 11.88x^2$.

   En la siguiente gráfica se representan las aproximaciones obtenidas sobre los datos. Ninguna de ellas, ni lineal básica ni polinómica, son capaces de ajustar la curva de forma precisa. Cuanto tenemos datos muy estructurados, como en este caso, una regresión global no suele producir buenos resultados.



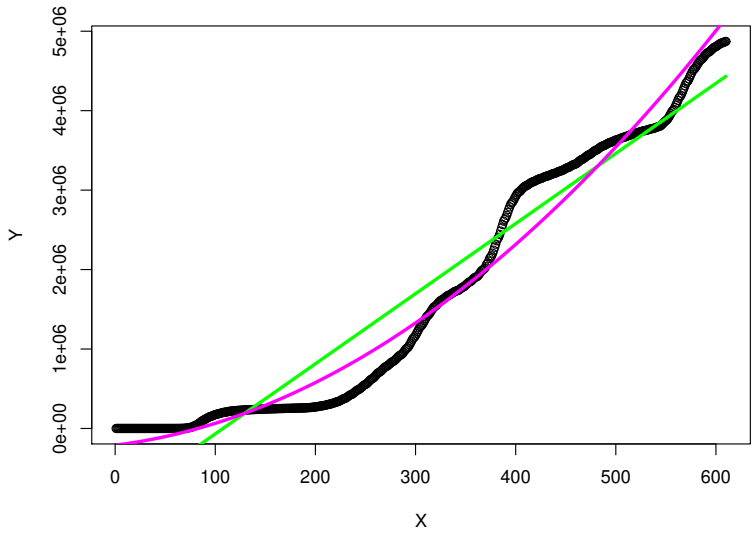
Regresión segmentada (rojo) sobre los datos de casos acumulados (negro).

4. Los puntos de corte entre segmentos consecutivos se pueden calcular fácilmente igualando sus expresiones. Si asumimos que uno es de la forma $\hat{y}_1 = c_{1,0} + c_{1,1}x$ y el otro es de la forma $\hat{y}_2 = c_{2,0} + c_{2,1}x$, tenemos que

$$\hat{y}_1 = \hat{y}_2 \quad \Rightarrow \quad c_{1,0} + c_{1,1}x = c_{2,0} + c_{2,1}x \quad \Rightarrow \quad x = \frac{c_{1,0} - c_{2,0}}{c_{2,1} - c_{1,1}}.$$

Con la expresión anterior, se obtienen los siguientes instantes de cambio de tendencia:

- Pto. de corte entre segmentos 1 y 2: $x = 71.56364$
- Pto. de corte entre segmentos 2 y 3: $x = 108.1001$
- Pto. de corte entre segmentos 3 y 4: $x = 227.6038$
- Pto. de corte entre segmentos 4 y 5: $x = 373.0229$
- Pto. de corte entre segmentos 5 y 6: $x = 401.3092$
- Pto. de corte entre segmentos 6 y 7: $x = 545.4047$



Regresión lineal básica (verde) y regresión polinómica (magenta) sobre los datos de casos acumulados (negro).

Evidentemente, los puntos de corte obtenidos están cerca de los extremos que hemos decidido para los segmentos. Cuando los datos son más dispersos, este punto es más complicado de determinar.

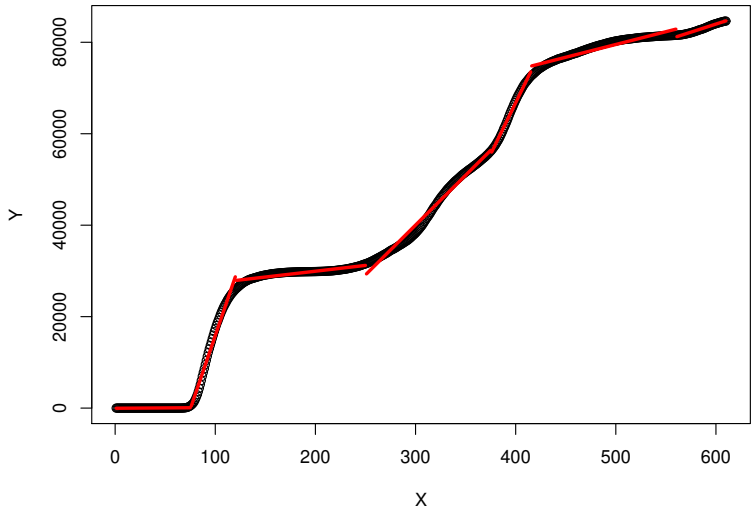
5. Repetimos los apartados anteriores para la curva de muertes acumuladas.

   De nuevo, hemos elegido 7 segmentos, cuyas fechas de inicio son son las que se corresponden con los días 1, 76, 121, 251, 376, 416 y 561, y sus fechas de fin se corresponden con los días 75, 120, 250, 375, 415, 560 y 610. Las aproximaciones para cada uno de los segmentos son:

   - Segmento 1: $\hat{y} = -37.32 + 1.52x$.
   - Segmento 2: $\hat{y} = -48414.89 + 642.72x$.
   - Segmento 3: $\hat{y} = 24849.47 + 25.42x$.
   - Segmento 4: $\hat{y} = -25367.63 + 217.97x$.
   - Segmento 5: $\hat{y} = -113437.85 + 450.55x$.
   - Segmento 6: $\hat{y} = 51584.27 + 55.86x$.

- Segmento 7: $\hat{y} = 41575.55 + 70.67x$.

En la siguiente gráfica se representan los segmentos obtenidos sobre los datos. Observamos que la regresión segmentada ajusta de forma óptima los datos.



Regresión segmentada (rojo) sobre los datos de muertes acumuladas (negro).

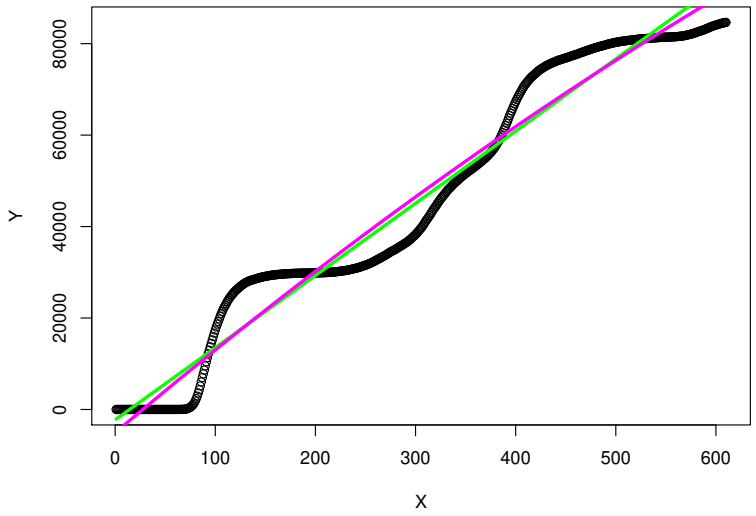
Las aproximaciones obtenidas para el apartado 3 son:

- Regresión lineal ($n = 1$): $\hat{y} = -2276.27 + 157.85x$.
- Regresión polinómica ($n = 2$): $\hat{y} = -5090.49 + 185.44x - 0.04515x^2$.

En la siguiente gráfica se representan las aproximaciones obtenidas sobre los datos. Ninguna de ellas, ni lineal básica ni polinómica, son capaces de ajustar la curva de forma precisa. Cuanto tenemos datos muy estructurados, como en este caso, una regresión global no suele producir buenos resultados. Además, en este caso, la regresión polinómica prácticamente no mejora a la regresión mediante una recta.

Para el apartado 4, se obtienen los siguientes instantes de cambio de tendencia:

- Pto. de corte entre segmentos 1 y 2: $x = 75.44792$



Regresión lineal básica (verde) y regresión polinómica (magenta) sobre los datos de muertes acumuladas (negro).

- Pto. de corte entre segmentos 2 y 3: $x = 118.6848$
- Pto. de corte entre segmentos 3 y 4: $x = 260.8003$
- Pto. de corte entre segmentos 4 y 5: $x = 378.6616$
- Pto. de corte entre segmentos 5 y 6: $x = 418.1079$
- Pto. de corte entre segmentos 6 y 7: $x = 676.001$

Evidentemente, los puntos de corte obtenidos están cerca de los extremos que hemos decidido para los segmentos. Cuando los datos son más dispersos, este punto es más complicado de determinar.

Código R

```
######### Ejercicio 2 #########
## Lectura de los datos
```

```
Y = myReadData_byDate('WHO-COVID-19-global-data-SPAIN.csv', '03/01/2020', '03/09/
    2021', 'Cumulative_cases')
#Y = myReadData_byDate('WHO-COVID-19-global-data-SPAIN.csv', '03/01/2020', '03/09/
    2021', 'Cumulative_deaths')
m = length(Y)
X = 1:m

# Representacion de los datos
plot(X, Y)

## Obtencion y representacion de cada uno de los segmentos (en este caso se
    utilizan 7 segmentos)
seg = 7
ini_segmentos = c(1, 81, 111, 221, 371, 401, 551)
fin_segmentos = c(80, 110, 220, 370, 400, 550, m)
#ini_segmentos = c(1, 76, 121, 251, 376, 416, 561)
#fin_segmentos = c(75, 120, 250, 375, 415, 560, m)
Cks = matrix(0, 2, seg)
for (i in 1:seg){

  s = ini_segmentos[i]:fin_segmentos[i]
  Xs = X[s]
  Ys = Y[s]
  Cks[, i] = mypolyfit(Xs, Ys, 1)
  lines(Xs, myeval(Xs, Cks[, i]), col='red', lwd=3)

}

## Comparacion de la regresion lineal con la regresion segmentada
plot(X, Y)
# Ajuste por regresion lineal (n=1) y representacion
Ck1 = mypolyfit(X, Y, 1)
lines(X, myeval(X, Ck1), col='green', lwd=3)

# Ajuste por regresion lineal polinomica de grado 2 (n=2) y representacion
Ck2 = mypolyfit(X, Y, 2)
lines(X, myeval(X, Ck2), col='magenta', lwd=3)

## Puntos de corte de las rectas
for (i in 1:seg-1){

  xs = (Cks[1,i] - Cks[1,i+1])/(Cks[2,i+1] - Cks[2,i])
  print(xs)

}
```

Criterios de corrección y puntuación de los apartados

Esta PEC tendrá un valor de **10 puntos** repartidos como sigue.

- Ejercicio 1. Este ejercicio tendrá un valor de **5 puntos**:
  - Tarea 1. La justificación de la elección de la técnica de linialización se valorará con **0.5 puntos** y el ajuste mediante la misma se valorará con **0.5 puntos**. **Total 1 punto**.
  - Tarea 2. El cálculo correcto de los coeficientes de la regresión lineal se valorará con **0.5 puntos**. **Total 1 punto**.
  - Tarea 3. La representación gráfica se valorará con **0.5 puntos** (0.25 puntos para cada modelo) y el comentario razonado se valorará con **0.5 puntos**. **Total 1 punto**.
  - Tarea 4. El cálculo correcto de los coeficientes de correlación se valorará con **0.5 puntos** (0.25 puntos para cada modelo) y el comentario razonado se valorará con **0.5 puntos**. **Total 1 punto**.
  - Tarea 5. La predicción del escenario futuro mediante ambos modelos vale **0.5 puntos** (0.25 puntos para cada modelo) y el comentario razonado vale **0.5 puntos**. **Total 1 punto**.
- Ejercicio 2. Este ejercicio tendrá un valor de **5 puntos**:
  - Tarea 1. La implementación correcta de la regresión segmentada se valorará con **1 punto**. **Total 1 punto**.
  - Tarea 2. La representación de la regresión segmentada se valorará con **0.5 puntos** y el comentario razonado se valorará con **0.5 puntos**. **Total 1 punto**.
  - Tarea 3. El ajuste mediante regresión lineal se valorará con **0.25 puntos** y mediante regresión lineal polinómica se valorará con **0.25 puntos**. La representación vale **0.25 puntos** y la comparación razonada de resultados vale **0.25 puntos**. **Total 1 punto**.
  - Tarea 4. El cálculo correcto de los puntos de corte vale **0.75 puntos** y el comentario razonado vale **0.25 puntos**. **Total 1 punto**.
  - Tarea 5. La repetición de los apartados anteriores con la curva de muertes acumuladas se valorará con **1 punto**. **Total 1 punto**.

Referencias

[1] Howard, J. P. (2017). Computational methods for numerical analysis with R. Nueva York: Chapman & Hall/CRC.

[2] Leitao Rodríguez, A.; Salvador Mancho, B.; Sancho Vinuesa, T. (2022). Aproximación de funciones y regresión. PID_00285421.2.