PEC 7 - Aproximación de funciones y regresión (I)

Fecha de entrega: 11/5/2022

Universitat Oberta de Catalunya

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.

Equivalente continuo de las curvas de evolución

La información y los datos que recibimos sobre la evolución de una pandemia se suelen presentar en tiempo discreto, es decir, nos proporcionan los valores de alguna magnitud (casos, hospitalizados, muertes) correspondientes a un momento concreto del tiempo (habitualmente cada día). Utilizando los datos en este formato se pueden realizar numerosos estudios estadísticos, como veremos en la siguiente PEC. Sin embargo, en muchas ocasiones, resulta muy interesante poder trabajar en tiempo continuo, esto es, con una función que asigna valores para cualquier instante de tiempo. Esto permite utilizar herramientas matemáticas clásicas tales como derivación o integración. En esta PEC construiremos el equivalente continuo de curvas relacionadas con la COVID-19 y lo emplearemos que realizar distintos análisis matemáticos.

1. Construcción del equivalente continuo de la curva de contagios (I)

Construir el equivalente continuo a la curva de nuevos contagios utilizando los polinomios de Legendre. Se utilizarán los datos de casos nuevos (etiqueta *New_cases*) entre el 01/12/2020 y el 31/03/2021, que se corresponden con la tercera ola de la pandemia en España. Incrementar progresivamente el grado del polinomio de Legendre hasta $n = 10$ y representar las distintas aproximaciones sobre los datos. ¿Que se observa en el ajuste? Justifica la respuesta.

Nótese que los polinomios de Legendre están formulados para trabajar el intervalo $[-1, 1]$. En este caso, los datos estarán en un intervalo genérico $[a, b]$, por lo que se debe aplicar una transformación adecuada para utilizar estos polinomios. Los valores de a y b serán el mínimo y el máximo de los datos, respectivamente.

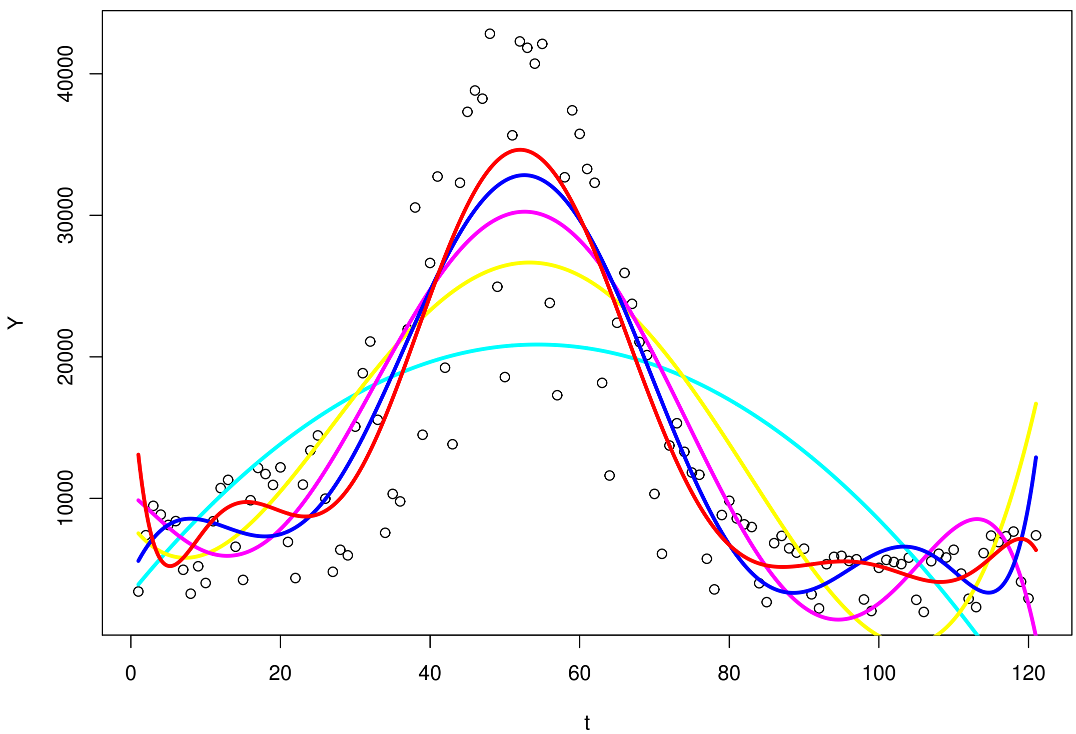
Solución

Obtenemos las siguientes aproximaciones:

- Para $n = 2$,
$$\hat{f}_L(x) = 13472.55L_0(x) - 4717.02L_1(x) - 14274.08L_2(x).$$
- Para $n = 4$,
$$\hat{f}_L(x) = 13472.55L_0(x) - 4717.02L_1(x) - 14274.08L_2(x) + 9299.31L_3(x) + 12915.31L_4(x).$$
- Para $n = 6$,
$$\begin{aligned}\hat{f}_L(x) = {} & 13472.55L_0(x) - 4717.02L_1(x) - 14274.08L_2(x) + 9299.31L_3(x) + 12915.31L_4(x) \\ & - 9406.2L_5(x) - 7073.06L_6(x).\end{aligned}$$
- Para $n = 8$,
$$\begin{aligned}\hat{f}_L(x) = {} & 13472.55L_0(x) - 4717.02L_1(x) - 14274.08L_2(x) + 9299.31L_3(x) + 12915.31L_4(x) \\ & - 9406.2L_5(x) - 7073.06L_6(x) + 8484.94L_7(x) + 4198.73L_8(x).\end{aligned}$$
- Para $n = 10$,
$$\begin{aligned}\hat{f}_L(x) = {} & 13472.55L_0(x) - 4717.02L_1(x) - 14274.08L_2(x) + 9299.31L_3(x) + 12915.31L_4(x) \\ & - 9406.2L_5(x) - 7073.06L_6(x) + 8484.94L_7(x) + 4198.73L_8(x) - 7033.56L_9(x) + 479.26L_{10}(x).\end{aligned}$$

Nótese que los coeficientes asociados a cada uno de los polinomios de Legendre son siempre los mismos, independientemente de número de polinomios que utilicemos.

Representando estas aproximaciones sobre los datos, obtenemos la siguiente gráfica. Se observa que, al aumentar el grado de los polinomios utilizados, la aproximación se ajusta mejor a los datos.



Aproximaciones mediante polinomios de Legendre para $n = 2$ (cian), $n = 4$ (amarillo), $n = 6$ (magenta), $n = 8$ (azul) y $n = 10$ (rojo).

Código R

```r

# Importamos la funcion de lectura de datos
source('Lectura_datos_por_fecha.R')

# Funcion que evalua los polinomios de Legendre hasta grado 'n' en 'x'
my_legendre = function(x, n){

  m = length(x)

  L = matrix(1, n+1, m)
  if (n > 0)
    L[2, ] = x

  if (n > 1)
    for (k in 1:(n-1))
      L[k+2, ] = ((2*k + 1)/(k + 1))*x*L[k+1, ] - (k/(k + 1))*L[k, ]

  return(L)
}

# Dados los coeficientes de Legendre, retorna la funcion aproximada evaluada en 'x'
myf_Legendre = function(t, ck){

  n = length(ck)-1

  x = 2*((t - min(t))/(max(t) - min(t))) - 1
  Lk = my_legendre(x, n)

  return(colSums(ck*Lk))
}

####### Ejercicio 1 #######

### Lectura de datos ###
# En 'Y' almacenamos los valores de casos nuevos y en 't' los dias
Y = myReadData_byDate('WHO-COVID-19-global-data-SPAIN.csv', '01/12/2020', '31/03/
    2021', 'New_cases')
m = length(Y)
t = 1:m
plot(t, Y)

####### Ejercicio 1 #######
## Cambio de variable
a = min(t)
b = max(t)
x = 2*((t - a)/(b - a)) - 1

N = c(2, 4, 6, 8, 10)
```

```
color = c('cyan', 'yellow', 'magenta', 'blue', 'red')
for(i in 1:length(N)) {

  n = N[i]

  ## Polinomios de Legendre
  Lk = my_legendre(x, n)
  Ak = 2/(2*(0:n) + 1)
  fLk = as.vector( Lk%*%Y )
  ckL = (2/(b - a))*fLk/Ak
  print(ckL)

  ## Representacion grafica
  hh = 0.01
  tt = seq(a, b, hh)
  fL = myf_Legendre(tt, ckL)
  lines(tt, fL, col=color[i], lwd=3)
}


```

2. Construcción del equivalente continuo de la curva de contagios (II)

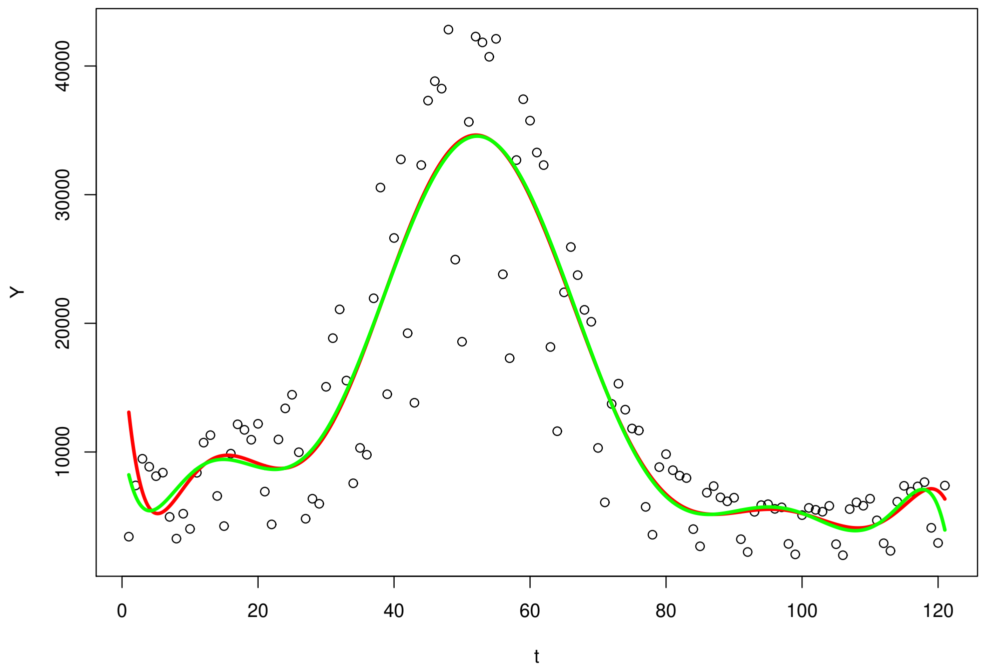
Realizar un ajuste polinómico sobre los mismos datos del apartado anterior y comparar con las aproximación obtenida con polinomios de Legendre mediante el coeficiente de determinación r^2. Representar gráficamente (sobre los datos) las dos aproximaciones (Legendre y polinómica) para $n = 10$. Aumentar el grado de ambas aproximaciones a $n = 15$. ¿Qué ocurre con el ajuste polinómico al aumentar el grado? Justifica la respuesta.

Solución

La aproximación mediante regresión lineal polinómica de grado $n = 10$ es

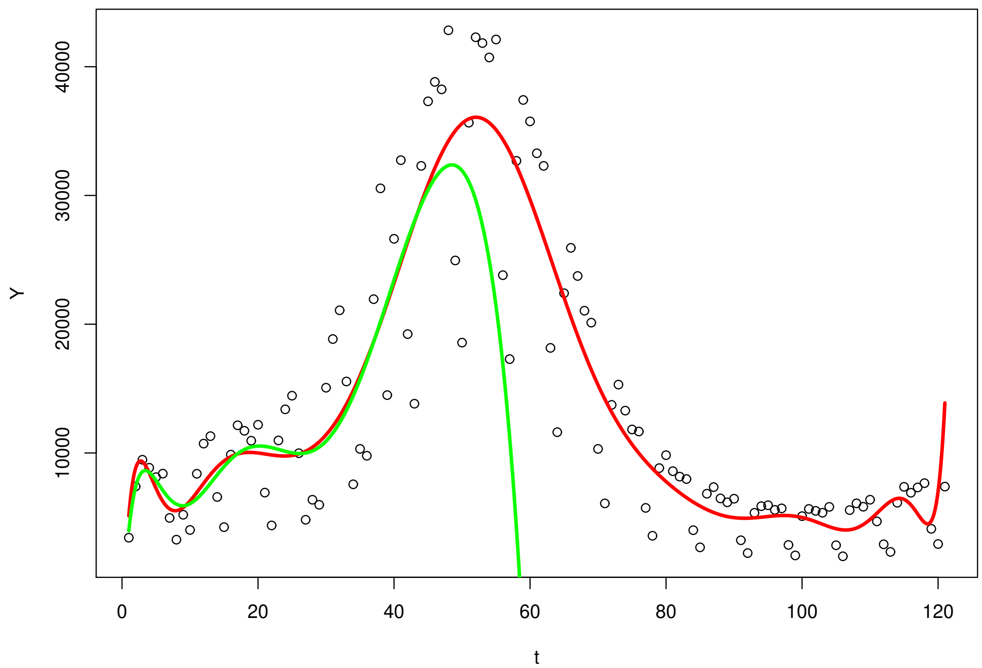
$$\begin{aligned}\hat{f}_P(x) &= 1.09 \cdot 10^4 - 3.37 \cdot 10^3 x + 6.90 \cdot 10^2 x^2 - 5.67 \cdot 10^1 x^3 + 2.31 x^4 - 5.04 \cdot 10^{-2} x^5 \\ &\quad + 5.84 \cdot 10^{-4} x^6 - 3.11 \cdot 10^{-6} x^7 - 2.15 \cdot 10^{-10} x^8 + 7.33 \cdot 10^{-11} x^9 - 2.22 \cdot 10^{-13} x^{10}.\end{aligned}$$

Representamos juntas las dos aproximaciones empleadas para $n = 10$. Observamos que la regresión polinómica se comporta de forma muy parecida a la aproximación por polinomios de Legendre. Para confirmarlo, calculamos los coeficientes de determinación, que resultan en 0.8018629 (Legendre) y 0.8048569 (ajuste polinómico).



Aproximaciones mediante polinomios de Legendre (rojo) y regresión lineal polinómica (verde) para $n = 10$.

Si aumentamos el grado de las aproximaciones hasta $n = 15$, obtenemos los ajustes representados en la siguiente gráfica. Observamos que la regresión lineal polinómica se comporta de forma inestable, tendiendo a menos infinito muy rápidamente. Este comportamiento erróneo puede deberse a una cuestión numérica (de representación computacional), que se podría mitigar aplicando un preprocesado a los datos. En cierto modo, los polinomios de Legrendre aplican un preprocesado de forma natural.



Aproximaciones mediante polinomios de Legendre (rojo) y regresión lineal polinómica (verde) para $n = 15$.

Código 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)

}

```

```
# Dados los coeficientes del ajuste polinomico, retorna la funcion aproximada evaluada en 'x'
myeval = function(x, c){

  f = 0
  for (i in 1:length(c)) {
    f = f + c[i]*x^(i-1)
  }
  return(f)
}

####### Ejercicio 2 #######
n=10

## Ajuste polinomico
lout = lsfit(myPhi(t, n), Y)
ckP = lout[[1]]

## Representacion grafica
hh = 0.01
tt = seq(a, b, hh)
fL = myf_Legendre(tt, ckL)
fP = myeval(tt, ckP)
plot(t, Y)
lines(tt, fL, col='red', lwd=3)
lines(tt, fP, col='green', lwd=3)

## Calculo del coeficiente de determinacion para Legendre y ajuste polinomico
St = sum( (Y - mean(Y))^2 )

Sr_L = sum( (Y - myf_Legendre(t, ckL))^2 )
Sr_P = sum( (Y - myeval(t, ckP))^2 )

r2_L = (St - Sr_L)/St
r2_P = (St - Sr_P)/St
print(r2_L)
print(r2_P)
```

3. Control de la evolución de las curvas

Uno de los mecanismos más importantes para controlar la evolución de las olas de una pandemia es detectar cuándo la pendiente de esta empieza a disminuir. Este hecho indica el momento en el que se empieza a “doblegar la curva”. Si se dispone de un equivalente continuo de la evolución temporal de los contagios (como los calculados en los apartados anteriores) se pueden utilizar técnicas clásicas de análisis matemático para determinar el instante en el que se produce el fenómeno de "doblegar la curva". En concreto, se sabe que la derivada de una función en un punto se corresponde con la pendiente de la tangente en ese punto. Por tanto podemos determinar que la curva empieza a doblegarse cuando la derivada pasa de crecer a decrecer.

Se pide calcular la derivada de alguna de las aproximaciones de los apartados anteriores (Legengre o polinómica) con grado $n = 10$. Aunque se puede calcular la derivada de forma analítica (exacta), se **deben** utilizar las técnicas de derivación numérica vistas en el curso. Representar la función derivada obtenida. ¿Qué observaciones se pueden extraer de ella? Justifica la respuesta.

Solución

Calculamos la derivada de la aproximación de Legendre para $n = 10$ mediante derivación numérica. Es decir,

$$\hat{f}_L'(x) \approx \frac{\hat{f}_L(x+h) - \hat{f}_L(x)}{h}.$$

Representamos la derivada numérica obtenida.

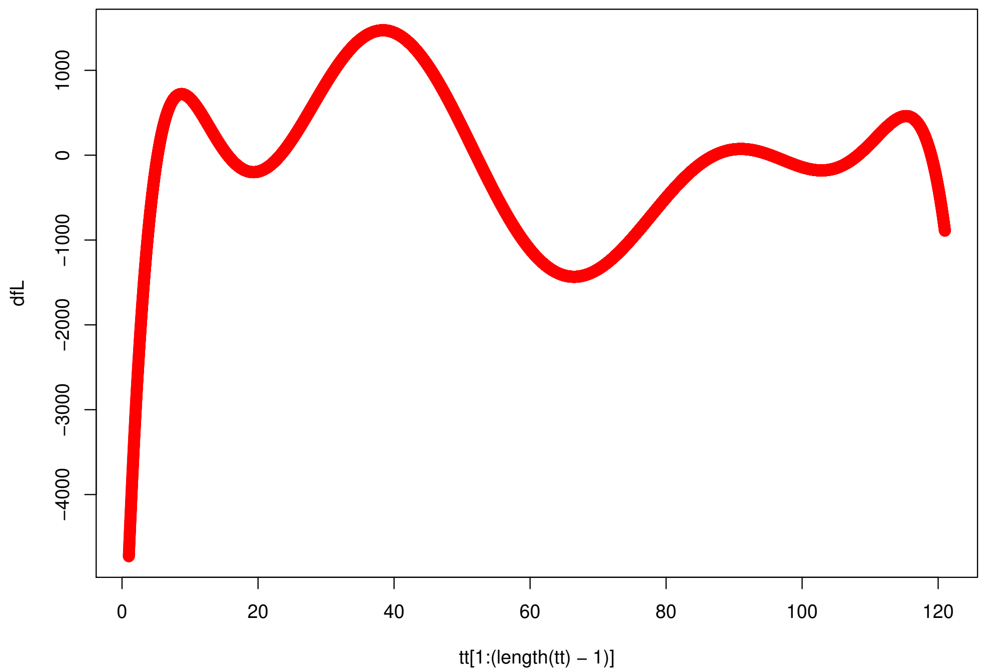
Si concentramos el análisis en el entorno de la parte importante de la curva (en la que el número de contagios es más alto), es decir, entre las abscisas $x = 20$ y $x = 80$, vemos que, en torno a $x = 45$ la derivada alcanza un máximo, indicando que la curva se empieza a doblegar (como se puede comprobar en las gráficas anteriores). Además, alrededor del punto $x = 55$, la derivada es cero, lo que indica que la curva ha llegado a su máximo, y comienza a bajar (de nuevo, se puede comprobar sobre las gráficas anteriores).

Como conclusión, podemos afirmar que disponer de un equivalente continuo es muy útil, ya que podemos detectar (mediante derivación) tanto el instante en el que la curva se empieza a doblegar como en el que alcanza su máximo.

Código R

```

# Implementacion de la derivacion numerica
my_diff = function(f, h){
  return(diff(f)/h)
}
```



Derivada numérica de la aproximaciones mediante polinomios de Legendre para $n = 10$.

```
####### Ejercicio 4 #######
# Calculamos la derivada de la aproximacion de Legendre y la representamos
hh = 0.01
tt = seq(a, b, hh)
fL = myf_Legendre(tt, ckL)
dfL = my_diff(fL, hh)
plot(tt[1:(length(tt)-1)], dfL, col='red', lwd=3)
```

Criterios de corrección y puntuación de los apartados

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

- Tarea 1: Se pide utilizar los polinomios de Legendre de distintos grados para construir el equivalente continuo de una curva de contagios. Se valorará con **2 puntos** la correcta construcción de la aproximación por polinomios de Legendre, con **1 punto** la representación gráfica y con **1 punto** la respuesta justificada a la pregunta. Total: **4 puntos**.
- Tarea 2: Se pide utilizar un ajuste polinómico para construir el equivalente continuo de una curva de contagios. Se valorará con **1 punto** la correcta utilización del ajuste polinómico, con **1 punto** la comparación en base al coeficiente de determinación, con con **1 punto** la representación gráfica y con **1 punto** la respuesta justificada a la pregunta. Total: **4 puntos**.
- Tarea 3: Se pide calcular la función derivada de la curva de contagios y representarla. Se valorará con **0.5 puntos** el cálculo de la derivada, con **0.5 puntos** su representación y con **1 punto** la respuesta justificada a la pregunta. Total **2 puntos**.

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.