Teoría de la aproximación: cuadrados mínimos

Juan Hirschmann - jhirschmann@fi.uba.ar

1. Introducción

1.1. Notación utilizada en el apunte

Con el propósito de evitar malentendidos, se explicitará la notación utilizada y su significado en el presente apartado. En primer lugar, se definirá al error cuadrático como:

$$S_r = \sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} \left(y_i - \sum_{j=0}^{m} C_j \phi_j(x_i) \right)^2 \quad \begin{cases} y_i : \text{Medición i-ésima} \\ C_j : \text{Coeficiente j-ésimo} \\ \phi_j : \text{Función elemental j-ésima} \end{cases} \tag{1}$$

Para aclarar la expresión, se considera que la función de ajuste es combinación lineal de los elementos $\phi_j(x_i)$ con términos C_j tal que minimizan el error cuadrático. Por su parte, los elementos $\phi_j(x_i)$ definen la base de funciones con las que se realizará el ajuste. Luego, el vector obtenido tras aplicar la función elemental $\phi_j(x_i)$ a cada variable x_i, se denotará ϕ_j. Por último, al minimizar ese error cuadrático se alcanzó la siguiente expresión:

$$\begin{pmatrix} < \phi_0; \phi_0 > & \dots & < \phi_0; \phi_m > \\ \vdots & \vdots & \vdots \\ < \phi_0; \phi_m > & \dots & < \phi_m; \phi_m > \end{pmatrix} \begin{pmatrix} C_0 \\ \vdots \\ C_m \end{pmatrix} = \begin{pmatrix} < Y; \phi_0 > \\ \vdots \\ < Y; \phi_m > \end{pmatrix} \begin{cases} <;> : \text{Operador producto interno canónico} \\ Y : \text{Vector que contiene a todas las observaciones} \end{cases} \tag{2}$$

1.2. Aplicación al ajuste no lineal

A lo largo del apunte se resolverá el ajuste para casos lineales, sin embargo, también es posible aplicar cuadrados mínimos para algunos casos en los que la función de ajuste no es combinación lineal de una base.

Un ejemplo de ello, puede ser el de obtener la amplitud y frecuencia angular del muestreo de una señal senoidal. En este caso, la función de ajuste será la siguiente:

$$A\text{sen}(\omega x) \tag{3}$$

Por lo tanto, su error cuadrático será:

$$\sum_{i=0}^{m} \left[A\text{sen}\left(\omega x_i\right) - y_i \right]^2 \tag{4}$$

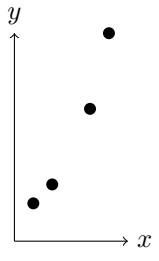
Luego, se deriva respecto a cada coeficiente para minimizar el error cuadrático:

$$\begin{cases} \frac{\partial}{\partial A} \sum_{i=0}^{m} \left[A\text{sen}\left(\omega x_i\right) - y_i \right]^2 = \sum_{i=0}^{m} 2 \left[A \sin\left(\omega x_i\right) - y_i \right] \sin\left(\omega x_i\right) = 0 \\ \\ \frac{\partial}{\partial \omega} \sum_{i=0}^{m} \left[A\text{sen}\left(\omega x_i\right) - y_i \right]^2 = \sum_{i=0}^{m} 2 \left[A \sin\left(\omega x_i\right) - y_i \right] A x_i \cos\left(\omega x_i\right) = 0 \end{cases} \tag{5}$$

Por último, se observa que las expresiones definen un sistema de ecuaciones no lineales. Este tipo de sistemas pueden no tener resolución analítica y se pueden resolver mediante alguno de los métodos estudiados en el curso

2. Ejemplo I: recta de regresión para una muestra reducida

Si bien en la práctica los ajustes se suelen realizar para mediciones numerosas, para familiarizarse con la mecánica del método es conveniente realizar un ajuste para pocas mediciones. Habiendo hecho mención de ello, se pide hallar la recta que mejor ajusta los siguientes datos:



x	y
1	2
2	5
4	7
5	11

Aunque la muestra, al ser tan reducida, no presenta un patrón claro, se ajustarán los datos mediante una recta. De esta forma, quedan definidas las funciones elementales sobre las cuales se basará el ajuste:

$$\phi_0(x) = 1, \quad \phi_1(x) = x \implies y_{ajuste} = C_0 + C_1 x \tag{6}$$

Luego, se construyen los vectores ϕ_0, ϕ_1 e Y:

$$\phi_0 = \begin{pmatrix} \phi_0(x_0) \\ \phi_0(x_1) \\ \phi_0(x_2) \\ \phi_0(x_3) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}; \quad \phi_1 = \begin{pmatrix} \phi_1(x_0) \\ \phi_1(x_1) \\ \phi_1(x_2) \\ \phi_1(x_3) \end{pmatrix} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 5 \end{pmatrix}, \quad Y = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 7 \\ 11 \end{pmatrix} \tag{7}$$

A continuación, se aplica la ecuación 2:

$$\begin{pmatrix} <\phi_0;\phi_0> & <\phi_0;\phi_1> \\ <\phi_0;\phi_1> & <\phi_1;\phi_1> \end{pmatrix} \begin{pmatrix} C_0 \\ C_1 \end{pmatrix} = \begin{pmatrix} <Y;\phi_0> \\ <Y;\phi_1> \end{pmatrix} \tag{8}$$

En donde:

$$<\phi_0;\phi_0> = 4, \quad <\phi_0;\phi_1> = 12, \quad <\phi_1;\phi_1> = 46, \quad <Y;\phi_0> = 25, \quad <Y;\phi_1> = 95 \tag{9}$$

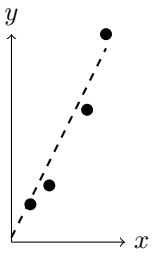
Por lo tanto, los coeficientes surgen de resolver el siguiente sistema de ecuaciones lineales:

$$\left(\begin{array}{cc|c} 4 & 12 & 25 \\ 12 & 46 & 95 \end{array}\right) \implies \begin{cases} C_0 = 0{,}25 \\ C_1 = 2 \end{cases} \tag{10}$$

Resolviendo el sistema, la recta de ajuste resulta:

$$2x + 0{,}25 \tag{11}$$

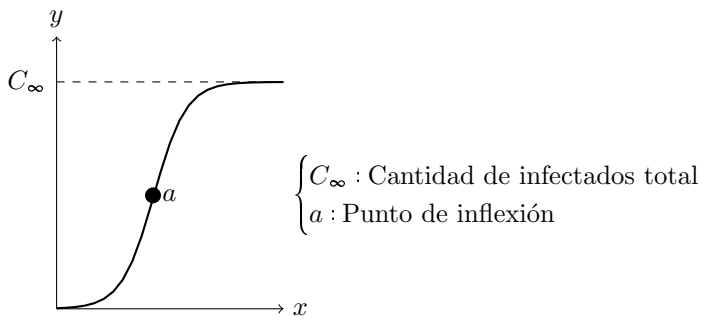
Así, se gráfica la función de ajuste de manera aproximada:



Comprender el funcionamiento a pequeña escala del ajuste de funciones permite aplicarlo a casos mas complejos y, por lo tanto, de mayor interés. En el próximo ejercicio se verá un ejemplo de ello.

3. Ejemplo II: evolución a corto y mediano plazo del coronavirus COVID-19 en Argentina

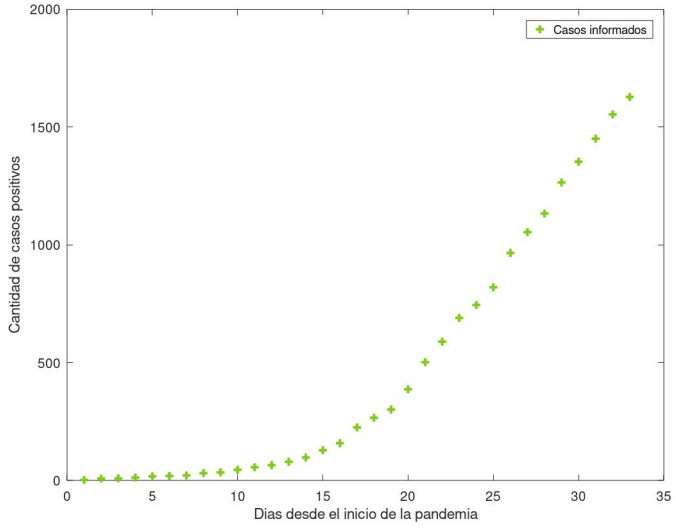
En el estudio de enfermedades infecto-contagiosas se conoce que, por su naturaleza, la propagación de una enfermedad de este tipo suele seguir una curva logística:



Analizando el gráfico, se pueden distinguir tres etapas. La primera, previa al punto de inflexión, se puede aproximar a una curva exponencial. La segunda, en un entorno del punto de inflexión, es similar a una lineal. Finalmente, la tercer etapa, se aproxima a una curva de saturación, $\alpha(1 - e^{-\beta x})$.

En otras palabras, los primeros días del brote la enfermedad crece a tasa exponencial ya que, al haber pocas personas infectadas, estas tienen una mayor probabilidad de interactuar con personas susceptibles y por lo tanto, mayor potencial de contagio. Luego, a medida que más personas se infectan o que el gobierno aplica medidas de mitigación, la cantidad de gente susceptible expuesta a la enfermedad baja y, por lo tanto, la tasa de contagios se vuelve constante. Finalmente, ya sea por el éxito de las medidas o porque todas las personas susceptibles se contagiaron, la enfermedad deja de propagarse.

Durante la pandemia de coronavirus del año 2020, localmente se registró un comportamiento similar al mencionado en la cantidad de infectados. Considerando el inicio de la pandemia en la Argentina el día 4 de Marzo del 2020, se tabularon y graficaron los casos informados por el ministerio de salud desde el 6 de marzo hasta el 6 de Abril del 2020. Utilizando consideraciones estadísticas, el dia 4 y 5 de marzo no fueron considerados en los cálculos.



Días desde el incio	Casos informados
2	8
3	9
4	12
5	17
6	19
7	21
8	31
9	34
10	45
11	56
12	65
13	79
14	97
15	128
16	158
17	225
18	266
19	301
20	387
21	502
22	589
23	690
24	745
25	820
26	966
27	1054
28	1133
29	1265
30	1353
31	1451
32	1554
33	1628

Como muestra el gráfico se propone realizar un ajuste por cuadrados mínimos de los datos tabulados utilizando dos aproximaciones: la primera, desde el dia 2 al 17 y la segunda desde el día 18 al 33.

3.1. Evolución a corto plazo del coronavirus

De observar el gráfico de contagios, se puede ver que hasta aproximadamente el día 17 la enfermedad crece exponencialmente. Por este motivo, se propone la siguiente función de ajuste

$$C_0 e^{C_1 x} = y \tag{12}$$

Se observa que la función de ajuste propuesta no se encuentra compuesta por una combinación lineal de funciones elementales. Sin embargo, al aplicar una transformación, se la puede llevar a una expresión lineal:

$$ln(C_0) + C_1 x = ln(y) \quad C_0' = ln(C_0) \tag{13}$$

De esta forma, el problema se reduce a calcular una regresión lineal para los datos transformados esto implica que $\phi_0 = 1$ y $\phi_1 = x$. Tras calcularla, se deben expresar los coeficientes recordando deshacer la transformación donde sea necesario. Por lo tanto, considerando que la recta de ajuste será combinación lineal de las funciones 1 y x, se pueden definir los siguientes

vectores:

$$\phi_0 = \begin{pmatrix} \phi_0(x_1) \\ \phi_0(x_2) \\ \phi_0(x_3) \\ \vdots \\ \phi_0(x_{16}) \end{pmatrix} == \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \phi_1 = \begin{pmatrix} \phi_1(x_1) \\ \phi_1(x_2) \\ \phi_1(x_3) \\ \vdots \\ \phi_1(x_{16}) \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ \vdots \\ 17 \end{pmatrix}, \quad Y = \begin{pmatrix} ln(8) \\ ln(9) \\ ln(12) \\ \vdots \\ ln(225) \end{pmatrix}, \quad \phi_{0,1}, Y \in \mathbb{R}^{16} \tag{14}$$

Replicando el desarrollo en el ejemplo anterior, se puede definir el siguiente SEL:

$$\begin{pmatrix} < \phi_0; \phi_0 > & < \phi_0; \phi_1 > \\ < \phi_0; \phi_1 > & < \phi_1; \phi_1 > \end{pmatrix} \begin{pmatrix} C_0' \\ C_1 \end{pmatrix} = \begin{pmatrix} < Y; \phi_0 > \\ < Y; \phi_1 > \end{pmatrix} \tag{15}$$

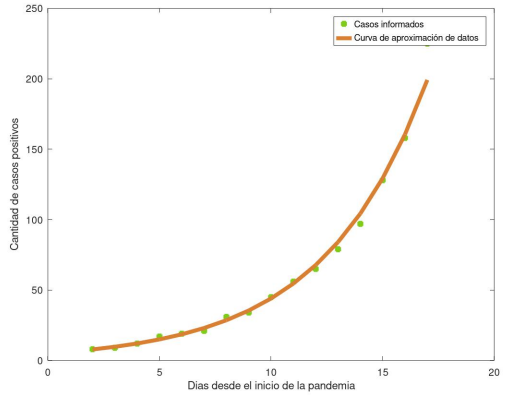
Tras realizar los productos escalares, se resuelve el sistema aplicando cualquier método estudiado:

$$\left(\begin{array}{cc|c} 16 & 152 & 58,825 \\ 152 & 1784 & 632,18 \end{array}\right) \Longrightarrow \begin{array}{c} C_0' = 1,6273 \\ C_1 = 0,21571 \end{array} \tag{16}$$

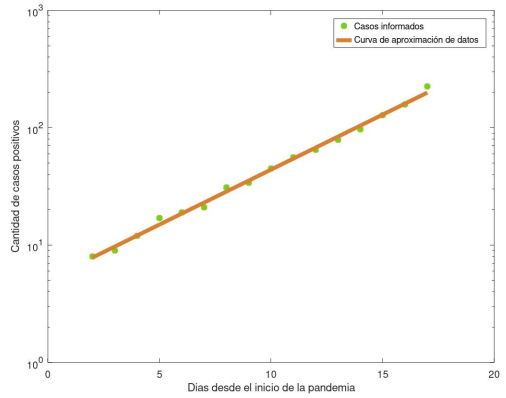
Por lo tanto, la función de ajuste resulta:

$$e^{1,6273} e^{0,21571 * x} \quad x \in [2, 17] \tag{17}$$

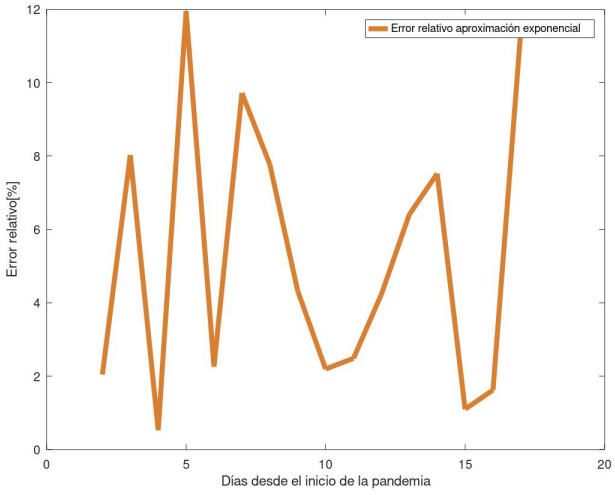
Con el fin de observar que tan fiable es el ajuste realizado, se graficó el ajuste en escala lineal y semilogarítmica, además, como parámetro adicional, se calculó el error relativo punto a punto:



(a) Escala lineal



(b) Escala semilogarítmica



(c) Error relativo al aproximar

Analizando el gráfico del error relativo, se puede ver que los datos se pueden considerar bien representados por la curva de ajuste dado que el error relativo es bajo punto a punto.

3.2. Evolución a plazo medio del coronavirus

A esta altura del ejercicio, se pueden alcanzar dos conclusiones posibles pero opuestas. La primera hipótesis es que el crecimiento sigue siendo exponencial, pero el testeo y la probabilidad de detección crecen linealmente. Esto implica que los datos no son una medida representativa de la realidad.

La segunda, es que el testeo es proporcional a la cantidad de infectados y, por lo tanto, el crecimiento de infectados es efectivamente lineal. Por el bien de todos, y de este ejercicio, se supone la segunda hipótesis correcta.

Habiendo mencionado ello, se considerará que el crecimiento de casos es lineal para el intervalo desde el día 18 al día 33 desde el comienzo del brote. Por lo tanto, se realizará el ajuste mediante una función lineal:

$$C_0 + C_1 x = y \tag{18}$$

De igual manera que en el ajuste para la parte exponencial de la curva, se definen los vectores, en este caso, las funciones elementales son las mismas pero no es necesario transformar el vector de mediciones datos:

$$\phi_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \phi_1 = \begin{pmatrix} 18 \\ 19 \\ 20 \\ \vdots \\ 33 \end{pmatrix}, \quad Y = \begin{pmatrix} 266 \\ 301 \\ 387 \\ \vdots \\ 1628 \end{pmatrix}, \quad \phi_{0,1}, Y \in \mathbb{R}^{33-17} \tag{19}$$

De manera análoga a la aproximación para el tramo anterior, se utiliza la expresión 2 y se define el SEL:

$$\begin{pmatrix} < \phi_0; \phi_0 > & < \phi_0; \phi_1 > \\ < \phi_0; \phi_1 > & < \phi_1; \phi_1 > \end{pmatrix} \begin{pmatrix} C_0 \\ C_1 \end{pmatrix} = \begin{pmatrix} < Y; \phi_0 > \\ < Y; \phi_1 > \end{pmatrix} \tag{20}$$

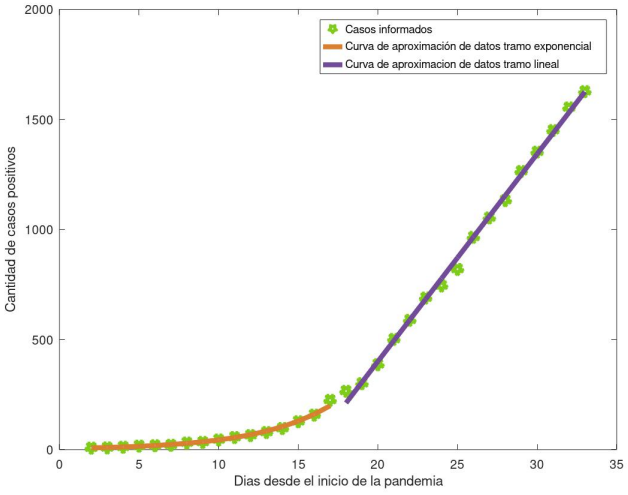
Luego, se realizan los productos internos:

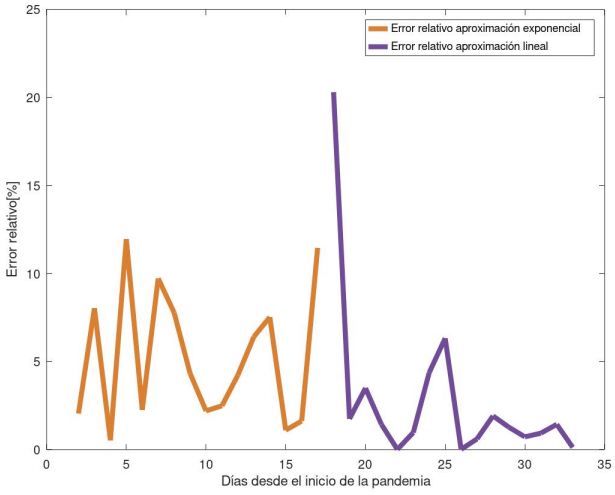
$$\left(\begin{array}{cc|c} 16 & 152 & 58,825 \\ 152 & 1784 & 632,18 \end{array}\right) \Longrightarrow \begin{array}{r} C_0 = -1484,825 \\ C_1 = 94,2676 \end{array} \tag{21}$$

De esta forma, la función que minimiza el error cuadrático para el tramo lineal de la curva de infectados resulta:

$$y = 94{,}2676x - 1484{,}825, \quad x \in [18, 33] \tag{22}$$

Utilizando ambas funciones de ajuste, se puede construir parte de la curva logística y expresar el error relativo para todos los puntos:





Como indican ambos gráficos, los datos parecen estar bien representados con ambos ajustes. Además, para la parte lineal de la curva, se revela una de las virtudes del ajuste por cuadrados mínimos: la medición correspondiente al día 18 se encuentra claramente por fuera de la tendencia lineal observada. Sin embargo, al ponderar todos los puntos por igual, la curva de ajuste se ve poco afectada por una medición en particular dada una muestra lo suficientemente grande.

4. Cuadro comparativo

Aproximación por cuadrados mínimos

Ventajas	Desventajas
Se obtiene una expresión simple para cualquier cantidad de datos	Cada punto tiene un error asociado
Un único punto no afecta desproporcionadamente al ajuste	La elección de una base de funciones elementales adecuada define la confiabilidad del ajuste
	El ajuste permite relacionar datos que pueden no estar relacionados entre sí