

AJUSTE LINEAL CON PUNTOS SINGULARES POR METODOS ITERATIVOS Y APLICACION BIOLOGICA

Por J. A. Carriedo*,
A. Gil** y
F. San Primitivo,*

INTRODUCCION

El análisis científico de los procesos y fenómenos biológicos precisa, frecuentemente, la aplicación de modelos matemáticos apropiados.

Cuando dos variables están correlacionadas, puede establecerse una función matemática que las relacione. De esta forma se obtiene un modelo matemático del proceso biológico. Entre los distintos modelos que en términos matemáticos expresan el fenómeno, trataremos de escoger el que se aproxime más a la realidad, es decir, el que mejor se ajuste.

El modelo matemático debe explicar la mayor cantidad posible de variación del carácter en estudio, a partir del conocimiento de otra u otras variables que se denominan explicativas.

Puede suceder que la relación entre la variable explicada o dependiente y la variable explicativa o independiente sea de tipo lineal, pero de distinta intensidad según el rango de valores que tome la variable independiente.

En estos casos, las observaciones podrán ajustarse a una línea quebrada con tantos tramos como número de respuestas distintas se tengan.

Un claro ejemplo de esta situación se tiene al observar los grados de desviación de la vertical con respecto al logaritmo en base 2 de la intensidad luminosa en Phycomyces (BERGMAN y col. 1973). BERGMAN reconoce un cierto grado de inseguridad en sus determinaciones, debido a que el ajuste a dos rectas, en este caso, se hace de forma arbitraria.

* Departamento de Genética y Mejora. Facultad de Veterinaria. León.

** Centro de Cálculo Electrónico. Consejo Superior de Investigaciones Científicas. Madrid.

An. Fac. Vet. León, 1978, 24, 139-148.

Nos proponemos con el presente trabajo obtener los métodos de ajuste de dicha quebrada, así como su aplicación a determinados procesos biológicos.

METODOS

1.º) Problema lineal general

Trataremos, en principio, el problema general para más tarde imponer una serie de restricciones que nos irán situando en problemas más específicos.

En este caso, partimos de una variable dependiente y , correlacionada con otra variable (en este caso independiente) x . Se trata de realizar el ajuste a partir de tantas funciones lineales o tramos como número de respuestas distintas se consideren.

El conjunto de estas funciones quedará definido en toda la recta real, constituyendo una función continua con tantos puntos singulares como intervalos a considerar menos uno.

En términos generales, si hemos de determinar un número n de intervalos en la recta real, estos intervalos quedarán definidos respecto a los correspondientes puntos singulares de la forma siguiente:

$$I_1 = (-\infty, x_m^{(1)}]$$

$$I_2 = [x_m^{(1)}, x_m^{(2)}]$$

$$I_3 = [x_m^{(2)}, x_m^{(3)}]$$

.

.

$$I_{n-1} = [x_m^{(n-2)}, x_m^{(n-1)}]$$

$$I_n = [x_m^{(n-1)}, +\infty)$$

Siendo $x_m^{(1)}$, $x_m^{(2)}$, ..., $x_m^{(n)}$ las abscisas de los puntos singulares.

El ajuste de las n funciones lineales se hace con las restricciones necesarias para que el conjunto de las n funciones constituya, a su vez, una función continua que relacione y con x . Esta función continua define, en realidad, la regresión de y sobre x .

Esta restricción consiste en que cada dos funciones consecutivas tomen el mismo valor en el punto singular que ambas determinan.

De esta manera, si la ecuación del primer tramo es $y = b_1x + a$, la ecuación del segundo tramo será $y = b_2(x - x_m^{(1)}) + b_1x_m^{(1)} + a$. La ecuación del tercer tramo será:

$$y = b_3(x - x_m^{(2)}) + b_2(x_m^{(2)} - x_m^{(1)}) + b_1x_m^{(1)} + a$$

y así sucesivamente, de forma que la ecuación del tramo l -enésimo, cuando $l > 2$, será:

$$y = b_l(x - x_m^{(l-1)}) + \sum_{i=2}^{l-1} b_i(x_m^{(i)} - x_m^{(i-1)}) + b_1x_m^{(1)} + a$$

Los valores de los parámetros a , b_1 , b_2 , ..., b_l , se obtendrán aplicando el criterio de mínimos cuadrados:

$$\min_{\nabla x} \sum (y - \hat{y})^2 = \min_{\nabla x} SC(\varepsilon)$$

donde y es el valor observado e \hat{y} el correspondiente valor estimado mediante la función.

Para que la suma de cuadrados de los errores

$$SC(\varepsilon) = \sum_{x_i \in I_1} (y_i - b_1x_i - a)^2 + \sum_{x_i \in I_2} (y_i - b_2x_i + b_1x_m^{(1)} - b_1x_m^{(1)} - a)^2 + \dots + \sum_{l=3}^n \sum_{x_i \in I_l} \left[(y_i - b_l(x_i - x_m^{(l-1)}) - \sum_{i=2}^{l-1} b_i(x_m^{(i)} - x_m^{(i-1)}) - b_1x_m^{(1)} - a)^2 \right]$$

sea mínima, las derivadas parciales de las sumas de cuadrados de los errores deben ser cero.

En definitiva se obtiene un sistema de $n + 1$ ecuaciones lineales con $n + 1$ incógnitas, cuya solución permite determinar el valor de los parámetros del problema.

Por otra parte, es posible calcular la varianza residual, simplemente dividiendo la suma de cuadrados de los errores entre la expresión: $N - (n + 1)$, donde N es el número total de observaciones.

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{\sum \varepsilon_i^2}{N - (n+1)}}$$

El coeficiente de determinación R^2 se calcula mediante la expresión

$$R^2 = 1 - \frac{(N - (n+1)) \hat{\sigma}_\varepsilon^2}{(N - 1) \hat{\sigma}_y^2}$$

donde $\hat{\sigma}_y$ es la varianza total de la variable y .

Según el método diseñado, es preciso definir las abscisas de los puntos singulares ($x_m^{(i)}$). A partir de estas constantes, se obtienen distintas soluciones, siguiendo una técnica de exploración sistemática en cada uno de los intervalos. De todas las soluciones obtenidas se selecciona la que maximiza R^2 (o minimiza SC) con lo que se termina una etapa. Al pasar a una nueva etapa de cálculo, es preciso definir de nuevo las abscisas de los puntos singulares que, en la práctica, corresponderán a los dos valores de $x_m^{(i)}$ más próximos (por defecto y por exceso) del considerado como óptimo en la etapa anterior.

Es preciso tener en cuenta que los intervalos donde se ensayan los puntos de separación son disjuntos y que sus amplitudes se van reduciendo en sucesivas etapas con lo que, siguiendo el criterio de maximizar el coeficiente de determinación, se puede conseguir que el error sea tan pequeño como se desee, al menos teóricamente.

Sin embargo, un número grande, e incluso moderado, de tramos, puede ocasionar, en la práctica, graves problemas de tiempo de procesamiento.

El método descrito sigue un sistema secuencial de elección de las abscisas de los puntos singulares. Estos valores de $x_m^{(i)}$ puede considerarse que constituyen una sucesión aritmética finita, cuya razón dependerá de la precisión que deseemos obtener.

El proceso puede ejecutarse de una sola vez. En este caso, para una determinada precisión, el número de etapas de cálculo precisas vendrán dadas por la expresión:

$$\sum_{i=1}^{n-1}$$

donde $m^{(i)}$ es el número de puntos $x_m^{(i)}$ probados en el paso y n el número de tramos.

Según esta expresión, cuanto mayor sea el número de tramos y/o el número de $x_m^{(i)}$ probados, mayor tiempo de procesamiento será necesario.

Este tiempo de procesamiento puede reducirse, realizando el proceso en varias fases. Dos serán suficientes en la mayoría de los casos. En la primera fase se reduce mucho el número de $x_m^{(i)}$ probados, buscando el intervalo en el cual se cumple que R^2 sea máximo. En la segunda fase, el intervalo del rango de valores a probar queda considerablemente reducido de tal forma que, por este sistema, puede lograrse la misma precisión, con menor tiempo de procesamiento.

2.º) Problemas específicos

A partir de este método general se han elaborado programas para solucionar los siguientes casos:

- a) Tres tramos sin restricciones.
- b) Dos tramos sin restricciones.
- c) Dos tramos con la restricción de que la pendiente del segundo tramo sea cero (plateau).
- d) N tramos por aplicación sucesiva del método de dos tramos sin restricciones.

a) Tres tramos sin restricciones

Puesto que es preciso determinar dos puntos singulares, $x_m^{(1)}$ y $x_m^{(2)}$, será menester especificar, en tarjeta, las amplitudes de los intervalos de variación de los puntos singulares, en cada paso, así como el número de puntos que se van a ensayar en dicho paso.

Dos, a lo sumo tres pasos, serán suficientes para obtener los dos puntos singulares con un error suficientemente pequeño, que minimice la varianza residual.

Este sistema podrá emplearse en problemas biológicos con dos cambios de tendencia o en problemas con un cambio de tendencia, para suavizar la variación de pendientes.

b) Dos tramos sin restricciones

En este caso tenemos un solo punto singular que determinar. Sin embargo, es preciso especificar la amplitud, del intervalo así como el número de puntos a ensayar en cada paso.

En aquellos casos en los que se desee una suavización en el lugar de cambio de pendientes, antes de la aplicación del correspondiente programa de tres tramos, será preciso utilizar el de dos tramos para conocer el rango de valores que pueden tomar los dos puntos singulares.

c) Dos tramos con «plateau»

En algunas ocasiones, a partir de un cierto valor de la variable independiente, la variable dependiente se hace constante. En otras ocasiones esta situación se intuye y por lo tanto, en al ajuste, deberá imponerse la restricción de que la pendiente de la función lineal del segundo tramo sea cero, a pesar de desconocer el punto a partir del cual se hace constante la variable dependiente.

Las ecuaciones de los dos tramos son en este caso:

$$\begin{aligned} y &= bx + a & \text{siendo } x \leq x_m \\ y &= b_1 x_m + a_1 & \text{siendo } x > x_m \end{aligned}$$

d) *Resolución del problema de n intervalos por aplicación sucesiva del método de dos tramos*

Para la resolución de este problema, el proceso se realizará por etapas, fijándose en cada una de ellas el punto singular obtenido.

En la primera etapa se calculará el valor de $x_m^{(1)}$ aplicando el programa de dos tramos sin restricciones. En la segunda etapa se fija el punto

$(y(x_m^{(1)}), x_m^{(1)})$ y se hace el ajuste mediante el programa de dos tramos con ligadura puntual, obteniéndose el valor del punto $x_m^{(2)}$.

Fijando ahora el punto $(y(x_m^{(2)}), x_m^{(2)})$, se repite el proceso y se determina el valor de $x_m^{(3)}$, continuando de esta misma manera hasta finalizar el número de tramos.

Para una mayor precisión, se podrá repetir el proceso tantas veces como se desee, dependiendo del error que estemos dispuestos a admitir. Siempre teniendo en cuenta los valores $x_m^{(i)}$ obtenidos en la etapa anterior, a partir de los cuales será factible obtener nuevos valores de $x_m^{(i)}$ con menor error.

Los programas de dos tramos se aplican a dos intervalos que, inicialmente, deben fijarse de forma intuitiva, por conocimientos anteriores del problema.

RESULTADOS Y DISCUSIÓN

Han sido realizados programas en Fortran IV para el ajuste de funciones continuas lineales compuestas de tres tramos, dos tramos, dos tramos con restricciones y para la resolución del problema general de n tramos por aplicación sucesiva del método de dos tramos.

Se obtiene los parámetros de dichas funciones siguiendo el criterio de mínimos cuadrados, es decir, minimizando la suma de cuadros del error o la varianza residual.

Mediante un proceso iterativo se llega a la función que minimiza la varianza residual, ensayando sucesivamente valores de puntos de separación entre los intervalos y escogiendo los que maximizan el coeficiente de determinación.

$$\max \left\{ R^2(x_m^{(1)}, x_m^{(2)}, \dots, x_m^{(n-1)}) \right\}$$

Estos programas han sido utilizados por diversos autores en investigaciones relacionadas con problemas biológicos a los que haremos referencia más adelante.

Problemas biológicos que podrían tratarse estadísticamente con nuestros métodos, han sido planteados por numerosos autores, algunos de los cuales citamos a continuación. Sin embargo, creemos que su método de cálculo es menos preciso que el nuestro, ya que en la mayoría de los casos el punto o puntos singulares se sitúan de forma arbitraria.

MEHREZ¹⁰ utiliza el método de dos tramos con «plateau» para determinar la mínima concentración de amoníaco que determina una máxima tasa de fermentación ruminal. El mismo modelo fue utilizado por BUTTERY³ al relacionar los gramos de proteína microbiana sintetizada, por cada 100 gramos de materia orgánica digestible, en el rumen de la oveja. SUTTER y ROFFLER¹¹ utilizan el mismo modelo en problemas biológicos similares.

Modelos de dos tramos lineales han sido planteados por WALKER^{12, 13} y KIRK⁸ en estudios sobre nutrición, BUIS² en estudios sobre curvas de saturación, MAULEON⁹ al estudiar la duración del anoestro port-partum en relación a la época del parto (modelo de tres tramos), e incluso FISCHER⁶ en otro tipo de estudios.

Otro problema biológico al que se han aplicado modelos de tramos es el estudio del comportamiento fototrópico del *Phycomyces* (BERGMAN y col.¹).

La mayoría de los autores revisados, estudian problemas biológicos cuyos datos podrían ajustarse a los modelos descritos por nosotros, sin embargo, el tratamiento estadístico utilizado es, en general, menos riguroso.

SATTER y ROFFLER¹¹ emplean para el ajuste de sus datos el modelo:

$$y = A + Bx + Cx_D + Dx_3 \quad \text{donde:}$$

$$x_D = 0 \text{ cuando } x < 8,5$$

$$x_D = 1 \text{ cuando } x > 8,5$$

$$x_3 = x_D x$$

con lo cual, la pendiente del primer tramo sería B y la del segundo B + D.

En los programas propuestos por nosotros va incluida la restricción $C + Dx_m = C + D 8,5 = 0$, que da mayor rigurosidad al modelo.

MEHREZ¹⁰ aplica un sistema muy similar al empleado por nosotros, observando que en sus experimentos la bondad de ajuste fue mejor con este sistema que cuando se trataba de realizar mediante una curva de tipo exponencial, $y = A + Be^{-Cx}$.

Nuestros métodos son superiores en cuanto a la precisión, ya que el sistema iterativo de cálculo permite obtener los resultados con la precisión que se deseé.

El resto de los autores consultados se limitan a calcular el punto de separación (punto singular) de los tramos de forma intuitiva, teniendo en cuenta, simplemente, su conocimiento biológico de la situación.

En general, mediante un análisis de varianza aproximado se podrá determinar si el modelo propuesto «a priori» es suficientemente preciso o si existe algún otro, de mayor complejidad, al cual los datos se ajusten significativamente mejor, en cuyo caso sería preciso rechazar el modelo que peor se ajusta.

La aplicación práctica del programa de dos tramos con «plateau», para la obtención de los valores $y(X_m)$ es tanto más precisa cuanto mayor número de puntos existan en el «plateau», es decir, en el último tramo, y cuanto más largo sea éste.

Este programa ha sido utilizado por GUADA⁷ en el tratamiento estadístico de siete experimentos similares, cuyos resultados finales diferían en la posición del punto singular. En estos experimentos pudo observarse que cuanto menor era el número de puntos situados en el «plateau» menor era la precisión. De todas formas, los valores de $y(X_m)$ se obtienen con un error menor que sus correspondientes valores X_m .

El programa de dos tramos con «plateau» ha sido también utilizado para la determinación del punto a partir del cual la capacidad del aparato digestivo deja de ser limitante en la ingestión energética, FERNÁNDEZ⁵. En este caso, con objeto de contrastar la hipótesis de que el modelo de dos tramos con «plateau» era el adecuado, se determinó, mediante un análisis de varianza aproximado, que no existían diferencias significativas entre la variación explicada por este modelo y el de dos tramos sin restricciones. Se eligió el modelo de dos tramos con «plateau» por concordar mejor con la hipótesis biológica emitida.

En cuanto a los programas para dos y tres tramos, la precisión con la que se calculan los puntos singulares dependerá de la brusquedad del cambio de pendiente. A mayor brusquedad de cambio, mayor precisión en las estimaciones.

Modelos de tres tramos han sido utilizados a partir de datos inéditos suministrados por ESLAVA⁴ y referentes a la respuesta de determinados mutantes de *Phycomyces* a la intensidad luminosa. A partir de estos datos hemos comprobado cómo el modelo de tres tramos viene a resolver el problema expuesto por BERGMAN¹ y explicado anteriormente.

En la aplicación del método de tres tramos a los datos anteriormente citados hemos podido comprobar que existen ocasiones en las cuales la mejor solución matemática no concuerda con la mejor solución biológica, de tal forma que al llevar la solución del modelo matemático al modelo biológico, es preciso utilizar la solución matemática contigua, muy semejante a la solución matemática óptima. Naturalmente ambas soluciones matemáticas están muy próximas y por supuesto no existen entre ellas diferencias significativas.

En aquellos casos en los que la diferencia de las pendientes de dos tramos consecutivos no resulte realmente tan brusca como la solución obtenida mediante el modelo de dos tramos sin restricciones permita deducir, podrá emplearse el programa de tres tramos con objeto de suavizar la pendiente.

En estos casos, podrá emplearse un polinomio en un determinado tramo e igualar, en el punto de separación, las derivadas primera, segunda, etc., con objeto de que el cambio sea menos brusco y el modelo matemático se ajuste mejor al biológico.

Por ejemplo, si en un tramo tenemos un polinomio de grado dos y en el segundo tramo una recta, es decir:

primer tramo — $y = a + bx + cx^2$ siendo $\forall x \leq X_m$

segundo tramo — $y = d + ex$ siendo $\forall x > X_m$,

se podrá imponer, además de la restricción de continuidad:

$a + bX_m + cX_m^2 = d + eX_m$, la igualdad de derivadas en dicho punto, es decir: $b + 2cX_m = e$, con lo que se obtiene una función continua y derivable, pero cuya función derivada tendrá singularidades.

Podrá obtenerse también una suavización del cambio de pendiente entre dos tramos, aumentando el número de tramos con un ajuste lineal.

Los programas descritos en este trabajo se encuentran en poder de los autores, en el Centro de Cálculo Electrónico del Consejo Superior de Investigaciones Científicas, en Madrid y en el Departamento de Genética y Mejora de la Facultad de Veterinaria de León.

RESUMEN

Se han realizado programas en Fortran IV para ajustar funciones compuestas de varias funciones lineales o tramos y sujetas a restricciones necesarias según el problema específico, mediante procesos iterativos.

La función compuesta por dichos tramos debe ser siempre continua, lo que implicará la igualdad de cada dos funciones lineales consecutivas en el punto singular de separación.

Además de estas restricciones que serán necesarias en todos los problemas planteados, se impondrán o no otras restricciones específicas del programa considerado.

Mediante un proceso iterativo se irán obteniendo sucesivamente los valores de los puntos singulares y de los parámetros siguiendo el criterio de mínimos cuadrados (LS).

Por último se expone la utilización práctica de los programas realizados así como su aplicación a numerosos procesos biológicos sobre los que han sido probados.

SUMMARY

Programmes in Fortran IV have been carried out with the purpose of fitting functions composed of several linear functions or stretches which are

subject to the necessary restrictions according to the problem, by means of an interative process.

The function composed of the aforementioned stretches must always be contuous, which will involve the equality of each consecutive two linear functions in the singular point of separation.

Besides these restrictions which will be necessary in all the problems posed, other additions will either be imposed or not, according to the problem under consideration.

By means of an interative process, the values of the singular points and parameters will gradually be obtained, following the criterion of least scuares (LS).

Finally we have set out the use of the programmes carried out in biological problems and their practical application.

AGRADECIMIENTOS

Agradecemos la colaboración prestada por don Pelegrín Zorrilla, del Centro de Cálculo Electrónico del C.S.I.C., para la realización del presente trabajo.

BIBLIOGRAFIA

- 1) BERGMAN, K., ESLAVA, A. P. y CERDA-OLMEDO, E. (1973).-Mutants of *Phycomyces* with Abnormal Phototropism. *Molec. gen. Genet.*, **123**: 1-16.
- 2) BUIS, R. C. (1972).-Iron saturation of different cattle Transferrin phenotypes. En *XIIth European Conference on Animal Blood Groups and Biochemical Polymorphisms*, 207-210.
- 3) BUTTERY, P. J. (1977).-Aspects of the biochemistry of rumen fermentation and their implication in ruminant productivity. En *Recent Advances in Animal Nutririon*. Havesing, W. y Lewis, D. (Editores). Butter Worths. London-Boston, 8-28.
- 4) ESLAVA, A. P.-Datos sin publicar.
- 5) FERNANDEZ, F.-Datos sin publicar.
- 6) FISCHER, K. S. y WILSON, G. L. (1975).-Studies of Grain Production in *Sorghum bicolor* (L. moench). III. The relative importance of assimilate supply, grain growth capacity and transport system. *Aust. J. Agric. Res.*, **26**: 11-23.
- 7) GUADA, J. A.-Datos sin publicar.
- 8) KIRK, R. D. y WALKER, D. M. (1976).-Plasma urea Nitrogen as an Indicator of Protein Quality. II. Relationships between Plasma Urea Nitrogen, various Urinary Nitrogen Constituents, and Protein Quality. *Aust. J. Agric. Res.*, **27**: 117-127.
- 9) MAULEON, P. y THIMONIER, J. (1972).-Aspectos fisiológicos de la reproducción ovina. *ITEA*, **6**: 95-116.
- 10) MEHREZ, A. Z., ORSXOR, E. R. y McDONALD, I. (1977).-Rates of rumen fermentation in relation to ammonia concentration. *Br. J. Nutr.*, **38**: 437-443.
- 11) SATTER, L. D. y ROFFLER, R. E. (1977).-Influence of nitrogen and carbohydrate inputs on rumen fermentation. En *Recent Advances in Animal Nutrition*. Havesign, W. y Lewis, D. (Editores). Butterworths London-Boston, 25-49.
- 12) WALKER, D. M. y KIRK, R. D. (1975).-The Utilization by Preruminant Lambs of Milk Replacers Containing Isolated Soya Bean Protein. *Aust. J. Agric. Res.*, **26**: 1.025-35.
- 13) WALKER, D. M. y KIRK, R. D. (1975).-The Utilization by Preruminant Lambs of Isolated Soya Bean Protein in Low Protein Milk Replacers. *Aust. J. Agric. Res.*, **26**: 1.037-52.