Actualizaciones en regresión: suavizando las relaciones

E. Sánchez-Cantalejo Ramírez / R. Ocaña-Riola
Escuela Andaluza de Salud Pública. Granada.

Correspondencia: Emilio Sánchez-Cantalejo Ramírez. Escuela Andaluza de Salud Pública. Apdo. 2070. Campus Universitario de Cartuja. 18080 Granada.

Recibido: 21 de julio de 1995
Aceptado: 21 de octubre de 1996

(An update in regression: smoothing relationships)

Resumen

Una metodología muy utilizada al analizar distintos tipos de problemas de salud se basa en los modelos de regresión: lineal, logística, etc.; estos modelos dependen de un conjunto de parámetros que hay que estimar a partir de los datos del estudio. Sin embargo, tienen el inconveniente de ser muy rígidos en el sentido de imponer, en ocasiones, relaciones demasiado estrictas entre la variable resultado y las predictoras. Los métodos de regresión no paramétrica presentan la ventaja de no establecer *a priori* ninguna restricción, permitiendo así que los datos nos indiquen la forma funcional apropiada. En este artículo se presentan algunos métodos modernos de regresión no paramétrica que además de su utilidad *per se* sirven de inestimable ayuda en el proceso diagnóstico de los métodos de regresión paramétrica. La disponibilidad actual del *software* necesario debe posibilitar su mayor utilización, lo que redundará en una mejor comprensión de los problemas de salud estudiados.

Palabras clave: Alisamiento. Regresión no paramétrica. Modelos aditivos generalizados.

Summary

A frequently used methodology for the analysis of different kinds of health problems is based on regression models: lineal, logistic, etc.; these models depend on a set of parameters that must be estimated from the data. However, they present the drawback of being very rigid since, occasionally, they impose overly strict relations between the variables. Non-parametric regression methods present the advantage of not establishing *a priori* restrictions, allowing the data to indicate us the appropriate functional form. In this paper several modern non-parametric regression methods are presented that in addition to their usefulness *per se* can prone to be of invaluable help in the diagnostic process for parametric regression methods. The current availability of the necessary software should contribute to their increased use which, in turn, will probably lead to an improved understanding of the health problems under study.

Key words: Smoothing. Non-parametric regression. Generalized additive models.

God has not decreed that all regressions should be linear.
R.G. Miller Jr.

Introducción

Bajo el nombre de modelos de regresión se incluyen un conjunto de técnicas que tratan de explicar cómo cambia una variable, la llamada variable dependiente o resultado, cuando cambian otra u otras variables, denominadas independientes o predictorias. Lo que caracteriza en principio a las distintas clases de modelos de regresión es la naturaleza de la variable dependiente; así, con variables continuas la clase de los modelos de regresión lineal es la más utilizada; con variables resultado dicotómicas, el método más popular en la investigación sanitaria es el modelo de regresión logística, etc.

Si no se especifica otra cosa, cuando se habla de regresión lineal se entiende que el método de estimación de los coeficientes del modelo es el de los mínimos cuadrados, es decir, las estimaciones de los coeficientes se calculan de tal forma que la suma de los cuadrados de los residuales, entendidos como la diferencia entre lo observado y lo predicho por el modelo, sea lo más pequeña posible. Este método de estimación se viene utilizando desde que a finales del siglo XVIII lo propusiera uno de los considerados más eminentes matemáticos de la historia, F. Gauss (1777-1855). La popularidad de esta forma de estimar se debe, aparte de las propiedades estadísticas deseables de sus estimadores, a la relativa simpleza del cálculo implicado, incluso en el caso multivariante. Para que se cumplan esas propiedades estadísticas de los estimadores es necesario que la distribución de los errores sea la normal, que la varianza de éstos sea constante, etc., condiciones que no siempre se dan. Por otra parte, la presencia de observaciones «raras» *(outliers)*

y de observaciones influyentes también plantea problemas a la hora de estimar los parámetros del modelo; finalmente, la llamada colinealidad, es decir, la dependencia lineal de una variable predictora en función de las restantes, también causa dificultades a la hora de estimar.

Aunque alguno de ellos se propuso hace mucho tiempo, ha sido en los últimos 30 años, con el advenimiento de los ordenadores, cuando se han empezado a utilizar otros metodos alternativos al modelo clásico basado en los mínimos cuadrados, con el ánimo de superar los problemas anteriores. Ejemplos de estos modelos son aquél que estima los coeficientes haciendo mínima la suma de los valores absolutos de los residuales *(least-absolute-deviation regression),* la M-regresión, que de alguna manera es una generalización de los métodos anteriores, la Ridge regresión, etc.[1] De cualquier modo, estas alternativas al modelo clásico son restrictivas, en el sentido de que siguen manteniendo la linealidad entre la respuesta y las variables predictoras, lo que muchas veces es una importante limitación.

A partir de finales de los años setenta, aparecen en las revistas estadísticas especializadas distintos trabajos donde se proponen unos nuevos métodos de regresión mucho más flexibles que los tradicionales, con el único coste de una mucho mayor complejidad de cálculo. El objetivo de este trabajo es presentar y discutir tales métodos así como sus aplicaciones en el ámbito de la salud pública. El resto del artículo se estructura como sigue: en primer lugar se distinguen los llamados métodos de regresión paramétricos de los no paramétricos; a continuación se introduce el concepto de alisamiento y se detalla uno de los procedimientos univariantes más populares de alisamiento, conocido como LOWESS. Por último, se presentan los modelos aditivos generalizados como versión multivariante de los métodos anteriores.

Regresión paramétrica *versus* no paramétrica

En general, un modelo de regresión trata de describir la relación entre una variable respuesta Y y una o varias variables predictoras $X = (X_1, X_2, \ldots, X_p)$; como quiera que para un mismo conjunto de valores de las predictoras los valores de la respuesta pueden ser múltiples, la curva de regresión se suele expresar de la forma

$$Y = f(X) + \varepsilon$$

donde f es la función de regresión y ε representa el error aleatorio; otra expresión alternativa del mismo modelo es

$$E(Y/X) = f(X)$$

donde $E(Y/X)$ denota el valor medio de la respuesta, dados unos ciertos valores de las variables predictoras. Para el caso de la regresión lineal simple, la función que liga a la variable predictora X con el valor medio de la respuesta es del tipo $f(X) = \beta_0 + \beta_1 X$, función que depende de β_0 y β_1, los llamados parámetros del modelo, que hay que estimar a partir de las observaciones (x_i, y_i) realizadas; por razones obvias, este modelo pertenece a la clase de modelos de regresión paramétrica.

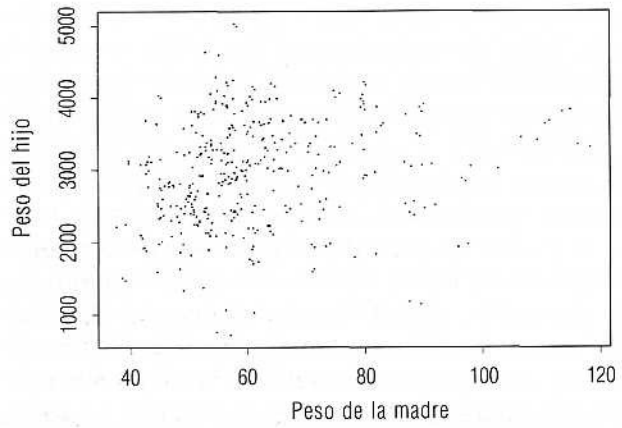
Si nuestros datos se ajustan de manera satisfactoria a este modelo, tendremos la suerte de poder explicar la relación entre la variable predictora y la respuesta de una manera muy sencilla; en efecto, cualquier lector con alguna cultura estadística conoce que el parámetro β_1 representa el cambio medio en la respuesta por unidad de aumento en la variable predictora. Obsérvese que lo que se está diciendo implícitamente es que siempre que la variable predictora X aumente en una unidad, el valor medio de la respuesta Y cambiará en una cantidad β_1, sea cual sea el valor de la variable predictora. Ésta es una condición demasiado fuerte en muchas ocasiones, pues el modelo puede ser válido en ciertos rangos de valores de X pero no en otros. Imagine el lector que se está estudiando la relación entre el peso del recién nacido, expresado en gramos, y el peso de la madre, expresado en kilogramos, y que se elige el modelo de regresión lineal para explicar esta relación; si a partir de los datos disponibles se consigue un valor de 10 como estimación del parámetro β_1, podríamos decir que los hijos de mujeres de un determinado peso tendrán, por término medio, 10 gramos más que los hijos de mujeres con un kg menos de peso, sea cual sea el peso de las madres; es decir, esa afirmación vale para madres delgadas, para madres con peso normal y para las que su peso no es tan normal. Pero, ¿no podría ocurrir que a mayor peso de la madre mayor peso del niño solamente en las madres delgadas y que para el resto el peso del recién nacido no dependiese del peso de la madre, manteniéndose constante? Más adelante volveremos sobre este ejemplo.

Para solventar estas dificultades, a partir de los años setenta se han propuesto distintas técnicas que se pueden englobar bajo el nombre de regresión no paramétrica, que tienen la ventaja de no proponer ninguna forma previa para la dependencia entre la variable resultado y las variables predictoras; por tanto, no tratan de estimar parámetros. Estos métodos no paramétricos son un intento de *dejar a los datos que nos muestren la forma funcional apropiada*[2]. Los modelos paramétricos, especialmente la regresión lineal, son métodos en ocasiones restrictivos, en el sentido de que imponen una relación entre las variables demasiado rígida.

Alisamiento

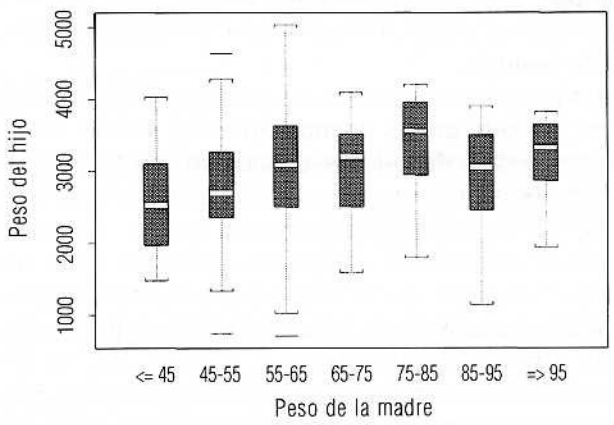
Antes de proponer cualquier modelo como candidato a explicar la relación entre dos variables es necesario representar gráficamente las parejas de valores de las dos variables, con objeto de evaluar la pertinencia de tal modelo. Sin embargo, este procedimiento es útil solamente cuando el número de parejas de valores representados es pequeño pues, de lo contrario, al ojo humano le es difícil captar la forma de la relación entre las dos variables. La figura 1 muestra el peso de 378 niños recién nacidos, junto con el de sus madres; esa nube de puntos difícilmente permite evaluar la forma funcional de la relación entre el peso del hijo y el de la madre.

Figura 1. Nube de puntos del peso de la madre, en kg, y el peso del hijo, en g



Un alisador es un procedimiento cuyo objetivo es mostrar la dependencia de una respuesta como función de una o varias variables predictoras; alisar una nube de puntos consiste en dar una estimación no paramétrica del valor medio de la respuesta para un valor dado de la predictora. Imaginemos en primer lugar que los datos provienen de un experimento en el que podemos fijar los valores X y disponer, por tanto, de varios valores de Y para cada valor de X; en esta situación, el valor medio de la respuesta en el valor x_i de la predictora se puede estimar calculando la media de los valores de Y correspondientes a ese valor x_i. Sin embargo, en la mayoría de las situaciones los datos provienen de estudios observacionales en donde no hay medidas repetidas para cada valor de X, y si las hay, éstas son poco numerosas, por lo que las estimaciones serán poco fiables, es decir, tendrán errores estándar grandes. Entonces la pregunta que se plantea es ¿qué hacer en este caso?

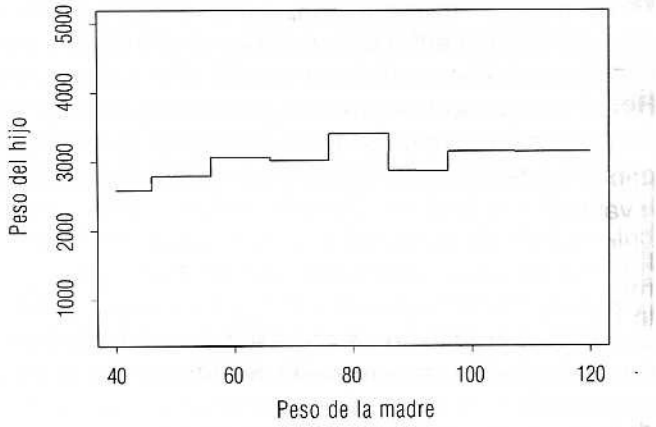
Figura 2. Cajas de los pesos de los hijos en los distintos grupos de madres



Una primera aproximación puede consistir en definir una serie de intervalos disjuntos para los valores de la variable predictora y ver cómo son los valores de Y en cada uno de tales intervalos; en la figura 2 aparece la distribución de los pesos de los niños, representada mediante una caja, para cada uno de los distintos intervalos de pesos de las madres. Esta representación gráfica nos da una primera pista del cambio de Y con el cambio de X.

También podríamos calcular la media de los valores de la variable predictora correspondientes a cada intervalo, estimando la media de Y en cada valor concreto de la predictora como la media de la respuesta correspondiente al intervalo en el que esté tal valor de X; este alisador se conoce con el nombre de regresograma. Como muestra la figura 3, éste no es el mejor mètodo de alisamiento, pues es una curva en escalera con un salto para cada intervalo.

Figura 3. Regresograma para la nube de puntos de la figura 1



Bajo el supuesto de que la función de regresión que queremos estimar no tiene discontinuidades, parece fácil admitir que a la hora de estimar $E(Y)$ en un valor x_i, aparte de contar con el valor o valores de Y en x_i, también pueden dar información los valores de la variable resultado correspondientes a valores de la variable predictora que estén próximos al valor x_i; dicho formalmente, para estimar la media de la respuesta en x_i se hará en base a los valores de Y correspondientes a un entorno de valores de x_i. Lo que queda por definir son los puntos que vamos a considerar como pertenecientes al entorno de x_i. Son dos las propuestas más usuales de definir los entornos; el método del k-entorno más próximo consiste en tomar como valores pertenecientes al entorno los $[k.n]$ valores de X más cercanos a x_i, siendo k, el llamado parámetro de alisamiento *(span)*, un número comprendido entre 0 y 1 con n representando el número de observaciones y $[k.n]$ el impar más cercano al producto $k.n$; es decir, los distintos entornos correspondientes a los distintos valores de X siempre contienen el mismo número de observaciones. El método de los núcleos consiste en elegir entornos de amplitud *(bandwith)* constante. Mientras que mediante el método de k-entorno más próximo el número de puntos en cada entorno es el mismo y por tanto, dependiendo de la densidad de puntos, la amplitud de los entornos cambiará, con el método de los núcleos la amplitud es constante pero el número de puntos dentro del entorno cambiará.

Dado un valor x_i y su entorno correspondiente, ¿cómo estimar la media de Y correspondiente a x_i? Una primera aproximación podría ser mediante la media de todos los valores de Y correspondientes a los valores X pertenecientes a su entorno; si y_i^s representa tal estimación, entonces

$$y_i^s = \frac{\sum_{j \in N_i} y_j}{[k.n]}$$

donde la suma se extiende a todos los elementos del entorno N_i correspondiente a x_i; este método de estimación es el llamado de las medias móviles.

Consideremos, por ejemplo, un parámetro de alisamiento $k = 0.1$, por lo que $[k.n] = 37$ y estimaremos el valor del peso medio de los niños de madres de 60 kg; para ello elegimos los 37 valores de peso más cercanos a 60, que son los que aparecen en la figura 4, entre las dos verticales rayadas. Entonces, la estimación del peso medio de los hijos de madres de 60 kg es la media de los pesos de los niños de las madres cuyos pesos figuran entre esas dos verticales; el peso medio de esos 37 niños es 2868,829 g.

Una alternativa al k-entorno más próximo es elegir como puntos constituyentes del mismo los $([k.n] - 1)/2$ puntos más próximos tanto por la derecha como por la izquierda, dando así lugar al método del k-entorno más próximo simétrico. Un ejemplo de alisador que utiliza el k-entorno más próximo simétrico es el llamado superalisador propuesto por Friedman[3], que utiliza un parámetro de alisamiento variable, lo que permite captar mejor las posibles curvaturas de la función a estimar. Segal[4] presenta una aplicación de este alisador al estudio del cambio del volumen expiratorio con el cambio de la edad.

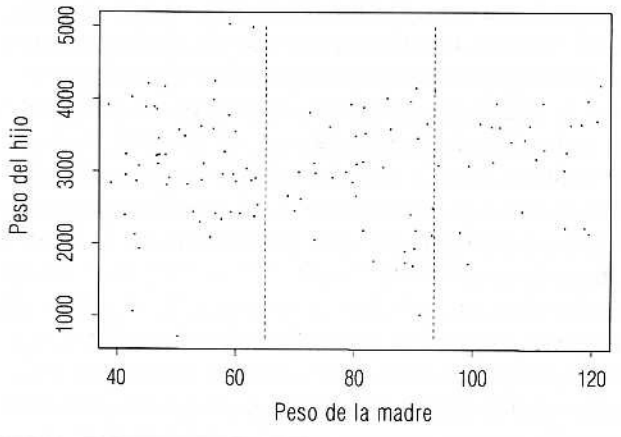
En cualquier de los métodos vistos hasta ahora para estimar la media del peso de los hijos de madres de 60 kg se está dando implícitamente la misma importancia a pesos de niños de madres con peso «cercano» a 60 kg que a pesos de niños con madres cuyo peso está «alejado» de 60 kg, todas ellas, por supuesto, dentro del entorno antes definido. Por tanto, la estimación que se acaba de hacer podría mejorarse asignando distintas importancias a los niños, dependiendo de lo similar que sea el peso de su madre al valor 60; es decir, las nuevas estimaciones se pueden definir por la expresión

$$y_i^s = \frac{\sum_{j \in N_i} w_j y_j}{\sum_{j \in N_i} w_j}$$

donde w_j es la importancia asignada a la observación j-ésima del intervalo.

¿Cómo asignar importancias a los distintos puntos de un entorno? Son muchas las funciones que se han propuesto para tal fin; entre ellas la función triangular, la de Epanechnikov, etc.[5], aunque estudios experimentales han demostrado la no excesiva influencia de la función que asigna las importancias sobre las estimaciones.

Figura 4. 0.1-entorno más próximo para el peso de 60 kg



LOWESS

Otra forma de estimar el valor medio de Y correspondiente a x_i es mediante lo que se denomina la regresión mínimo-cuadrática local que está basada en una idea bastante simple: se trata de calcular la recta de regresión lineal, estimada por el método de los mínimos cuadrados, a partir solamente de los puntos del entorno considerado y, a través de ella, estimar el valor correspondiente a x_i mediante la expresión

$$y_i^s = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

donde $\hat{\beta}_0$, $\hat{\beta}_1$ son los estimadores de los parámetros de la recta; así, para el intervalo considerado, la recta estimada es $y^s = 13588{,}73\text{-}177{,}97\ X$, por lo que la estimación del peso medio de los niños de madres con 60 kg, según la regresión local, es 13588,73-177,97(60) = 2910,53. Aunque a primera vista este método parezca un poco raro no hay que olvidar que mediante tal regresión lineal lo que se está estimando es también un valor medio.

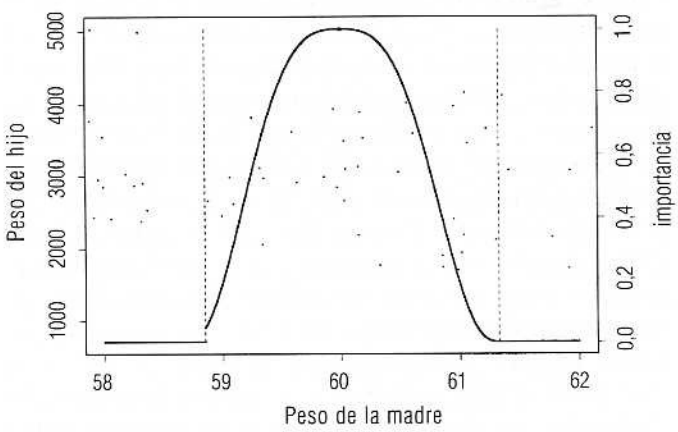
Cleveland[6] propuso un método de alisamiento basado en el método del *k*-entorno más próximo y estimando mediante regresión mínimo-cuadrática local ponderada (*LOcally WEighted regression Scatterplot Smoothing,* LOWESS); para la ponderación utilizó la llamada función tricubo, función definida como sigue:

$$w_j = \begin{cases} \left(1 - \left(\dfrac{|x_i - x_j|}{d_i}\right)^3\right)^3 & \text{si } |x_i - x_j| \le d_i \\ 0 & \text{en otro caso} \end{cases}$$

donde x_i es el valor de la predictora en donde se quiere estimar el valor medio de la variable resultado, x_j es cualquier valor perteneciente al entorno considerado y d_i es la máxima distancia de x_i a cualquier x_j. De estas consideraciones se deduce que el cociente $|x_i - x_j| / d_i$ está comprendido entre 0 y 1; por tanto, las importancias w_j son también número mayores o iguales a cero y menores o iguales a uno; fuera del intervalo, la función tricubo vale cero, es decir, los pesos de los hijos de madres cuyo peso esté fuera del intervalo no se consideran.

De la definición de la función tricubo se colige que ésta toma el valor más grande, la unidad, precisamente en el valor $x_j = x_i$, lo que se traduce diciendo que a la hora de estimar el peso medio de hijos de madres de 60 kg se da más importancia a los pesos de hijos de madres de 60 kg que a los pesos de los hijos de las otras madres del intervalo; cuanto más se aleje el peso de la madre de 60 kg, menos importancia tendrán los pesos de sus hijos. Debemos señalar, por último, que la función tricubo es simétrica respecto a la vertical trazada por el punto donde se hace

Figura 5. Representación gráfica de la función tricubo en el entorno de 60 kg



la estimación, en este caso 60, lo que implica que igual importancia tendrá el peso de un hijo de una madre con 59,5 kg que el de una con 60,5 kg. En la figura 5 aparece la representación gráfica de la función tricubo para el intervalo considerado; asignando a los 37 pesos de los niños del intervalo las importancias derivadas de esta función podemos calcular, mediante regresión lineal ponderada, una estimación del peso medio de hijos de madres con 60 kg; para nuestro ejemplo tal estimación es de 3030,78 g.

Ya que no es infrecuente la presencia de observaciones extremas *(outliers),* entendidas éstas como aquéllas en las que el valor de la respuesta es muy distinto de los valores de la respuesta para las observaciones próximas a ella en términos de la variable predictora, Cleveland propuso una estrategia para protegerse contra su presencia, es decir, dar robustez a las estimaciones, mediante la regresión mínimo-cuadrática ponderada iterativa. Este proceso se puede describir como sigue: una vez hechas las estimaciones como se acaba de indicar, se definen los residuales:

$$r_i = y_i - y_i^s$$

con lo que las observaciones extremas vendrán identificadas por tener residuales grandes en valor absoluto; se definen a continuación un nuevo conjunto de importancias, las importancias de robustez v_j, mediante la denominada función bicuadrado,

$$v_j = \begin{cases} \left(1 - \left(\dfrac{r_i}{6m}\right)^2\right)^2 & \text{si } \left|\dfrac{r_i}{6m}\right| < 1 \\ 0 & \text{en otro caso} \end{cases}$$

donde

$$m = \text{mediana } |r_j|$$

de tal manera que las observaciones correspondientes a residuales próximos a cero tendrán valores v_i altos y las observaciones con residuales grandes tendrán importancias de robustez pequeñas.

Con las importancias de regresión w_i asignadas del entorno y estas últimas de robustez v_i se pueden formar unas nuevas importancias mediante su producto $w'_i = w_i \cdot v_i$, que son las que se utilizarán como ponderación en una nueva regresión local ponderada. Obsérvese que la importancia asignada en este nuevo ajuste a la observación x_i cuando se está estimando el valor de la función en x_j será próximo a cero si la observación x_i está lejada de x_j y/o el valor y_i es una observación rara. Cleveland recomienda hacer este procedimiento de robustez dos veces.

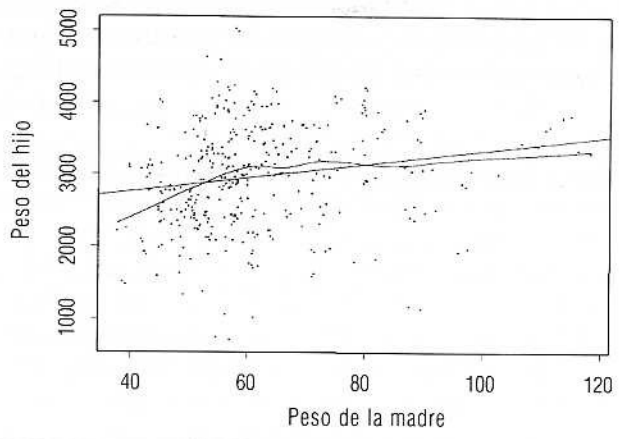
Sea cual sea el método de ponderación, lo que se acaba de describir para el peso de 60 kg habría que realizarlo para todos los pesos de las madres, por lo que obtendríamos una estimación del peso de los hijos para cualquier valor observado de peso de la madre; interpolando entre tales estimaciones se obtiene la estimación de la curva de regresión que venimos persiguiendo.

En la figura 6 se muestra el alisado de los pesos de las madres y de sus hijos mediante LOWESS con un parámetro de alisamiento de 2/3 y con dos iteraciones para el procedimiento de robustez; en el mismo gráfico también aparece la recta de regresión mínimo cuadrática cuya ecuación es:

$$\hat{y} = 2384{,}09 + 9{,}42X$$

por lo que según ésta, por cada aumento de un kg en el peso de la madre, el peso del hijo aumentará, por término medio, 9,42 g y esto es válido sea cual sea el peso de la madre (he aquí la rigidez del modelo lineal que antes de comentó). Sin embargo, la curva ajustada mediante el procedimiento de regresión no paramétrica muestra que en madres con peso inferior a 60 kg el peso del niño aumenta de forma lineal con el aumento del peso de la madre; para madres con más de 60 kg el peso de sus hijos permanece prácticamente constante. La relación que muestra la función alisada entre el peso de los niños y el de sus madres tiene más plausibilidad biológica que la de la recta de mínimos cuadrados. Greenland[7] utiliza este procedimiento de alisado para evaluar el cambio en el riesgo de infarto de miocardio con el consumo de café. Bush[8] también utiliza LOWESS para estudiar, en enfermos transplantados de médula ósea, la relación entre calidad de vida y tiempo transcurrido desde el transplante.

Figura 6. Curva de regresión estimada por el procedimiento LOWESS y recta de mínimos cuadrados



Elección del grado de alisamiento

Hasta ahora se ha hablado del parámetro de alisamiento o de la amplitud del entorno, pero no se han discutido criterios para su elección. El problema de la selección del parámetro de alisamiento es similar al de la selección de variables en el caso de la regresión múltiple. Si se elige como parámetro de alisamiento el valor $1/n$, significa que en cada entorno entrará un solo punto, por lo que la estimación en x_i será precisamente el valor observado y_i y no existirá ningún sesgo a la hora de estimar la función; por otra parte, ya que este procedimiento reproduce exactamente las observaciones realizadas, la función estimada será muy dentada, lo que implica que tendrá una gran varianza en caso de muestreo repetido. En el otro extremo, cuando se elija un parámetro de alisamiento o una amplitud de entorno muy grande la curva será muy alisada, con lo que el sesgo aumentará y la varianza disminuirá (el llamado *trade-off between bias and variance*).

Una forma de elegir el parámetro de alisamiento es mediante la inspección visual de las curvas estimadas en función de distintos valores de este parámetro. Para superar la subjetividad de este criterio se han propuesto distintas estrategias. El principio que las inspira es elegir el parámetro de alisamiento de tal manera que la suma de cuadrados de los residuales sea lo más pequeña posible; lo mismo que en la regresión múltiple, este criterio lleva a incluir en el modelo a todas las variables medidas, lo que en nuestro caso lleva a amplitudes cero o a parámetros de alisamiento igual a $1/n$. Para evitar esta dificultad se ha propuesto la técnica de validación cruzada local *(local cross-validation);* si k representa el parámetro de alisamiento, se trata de elegir su valor tal que minimice a la suma

$$\sum_{i=1}^{n} [y_i - y_{(i)}^{s}(x_i/k)]^2$$

donde $y_{(i)}^{s}(x_i/k)$ es la estimación local de Y correspondiente a x_i sin considerar el valor y_i; el superalisador uti-

liza una versión de esta estrategia. Otros criterios de elección del entorno caen fuera del planteamiento de este artículo.

Otros modelos de regresión no paramétrica. Los modelos aditivos generalizados

Un método de alisamiento bastante extendido es la llamada regresión por *splines*[9], que puede considerarse como un paso intermedio entre la regresión polinómica clásica y los métodos más modernos de regresión no paramétrica. Se trata de una regresión mediante polinomios a trozos *(piecewise)* que pueden tener formas muy variadas; el tipo de *spline* más utilizado es el de grado tres. Ejemplos de esta metodología aplicada al campo de la salud se pueden encontrar en Harrell[10] y en Durrleman[11]. Royston[12] propone un método de regresión basado en una familia de polinomios, los polinomios fraccionales, de los que los polinomios convencionales son un caso particular; los autores ejemplifican esta metodología aplicada al modelo lineal, al logístico y al de riesgos proporcionales con datos clínicos. Tanto los *splines* de orden tres como los polinomios fraccionales tienen la ventaja de poder ser realizados con los paquetes estadísticos que tengan implementada la regresión lineal.

Hasta ahora se han comentado diversos métodos de regresión no paramétrica para el caso de una sola variable predictora. La versión multivariante del procedimiento de Cleveland considera entornos *p*-dimensionales, siendo *p* el número de predictoras, basados en la distancia euclídea de los datos una vez normalizados. Es la versión no paramétrica del modelo:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

aunque presenta dos graves inconvenientes. El primero es el llamado «problema de la dimensionalidad»: el número de puntos de los entornos decrece con el aumento de *p*, por lo que si no se dispone de bases de datos muy grandes, el número de observaciones en cada entorno será pequeño y las estimaciones resultantes serán poco fiables; el segundo inconveniente es la imposibilidad de evaluar el efecto de cada una de las variables, una cuestión de mucho interés.

Los modelos aditivos constituyen un intento de superar estas dificultades. Conservan el carácter aditivo del modelo lineal, pero no son tan restrictivos. Un modelo aditivo es de la forma

$$Y = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) + \varepsilon$$

donde las f_i son funciones a estimar no paramétricamente mediante un proceso iterativo basado en el LOWESS o en los *splines;* evidentemente, cuando $f_i(X_i) = \beta_i X_i$, el modelo aditivo coincide con el modelo lineal. De forma similar a como McCullagh[13] extendió el modelo lineal a la clase de los modelos lineales generalizados, Hastie[14] generalizó esta metodología más allá del modelo aditivo a lo que se conoce como modelos aditivos generalizados para posibilitar otras distribuciones distintas a la normal, como es el caso de la distribución binomial y la distribución de Poisson. Estos modelos permiten que algunos de los componentes sean de carácter paramétrico y que algunas variables predictoras sean categóricas. El modelo logístico aditivo

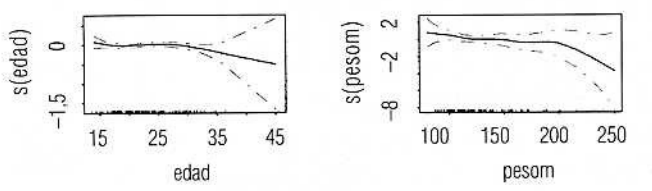
$$\log \frac{P}{1-P} = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$

es un caso particular de modelo aditivo generalizado.

Como aplicación práctica de este modelo, se consideran a continuación los datos sobre bajo peso al nacer analizados por Hosmer[15]. Se trata de una base de datos de 189 recién nacidos considerados como de bajo peso = 1 si pesaron menos de 2.500 g al nacer y 0 en el caso contrario. También se tiene en cuenta características de la madre como la edad, el peso (en libras), la raza, el hábito de fumar, el número de partos prematuros previos, la historia de hipertensión, la presencia de irritabilidad uterina y el número de visitas al médico en el primer trimestre de embarazo. Todas las variables, excepto el peso y la edad de la madre, son tratadas como categóricas; la cuestión que se plantea es cómo tratar estas dos variables continuas. En el modelo que estos autores presentan (tabla 4.8), tratan al peso de la madre como dicotómica, diferenciando, por una parte, las madres por debajo del primer cuartil y el resto, por otra. ¿Podría, sin embargo, evitar el inconveniente que supone la dicotomización de esta variable continua?

Ajustando el modelo logístico aditivo con las mismas variables predictoras, se pueden obtener las dos representaciones gráficas que aparecen en la figura 7; en ellas se muestran, con trazo continuo, las estimaciones de la forma funcional de las relaciones de la edad y el peso con el logit de la probabilidad de bajo peso al nacer; las líneas rayadas presentan el valor estimado ± dos errores estándar. Salvando los extremos dere-

Figura 7. Alisamiento de las variables edad y peso de la madre



chos de las dos relaciones estimadas, parece que no hay evidencias de no linealidad en tales relaciones, por lo que las variblaes edad y peso de la madre se pueden tratar como tales, no siendo necesaria ningún tipo de transformación. Las estimaciones realizadas para valores grandes de edad y de peso son poco fiables, lo que queda en evidencia por la amplitud de los intervalos en tales regiones. En definitiva, no es necesario dicotomizar el peso de la madre.

Ejemplos de la utilización de este modelo logístico se pueden encontrar en Herman[16], que estudia la relación entre la muerte neonatal, la edad y el tamaño del niño y en Hastie[17]. Con alguna modificación en el método de estimación, Hastie[18] generaliza el modelo de Cox en la forma

$$\lambda(t) = \lambda_0(t) \exp [f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)]$$

Por último, Schwartz[19] hace otra aplicación de esta metodología utilizando la versión no paramétrica del modelo de regresión de Poisson con el objetivo de estudiar la asociación entre el número de ingresos hospitalarios diarios por neumonía y la concentración ambiental de partículas de diámetro inferior a 10 µm.

Conclusiones

Aunque los modelos de regresión cuyo componente sistemático es una combinación lineal de las variables predictoras son muy populares en la literatura sanitaria como herramienta para el estudio de las relaciones entre dos o más variables, dichos modelos adolecen de falta de flexibilidad. Los métodos de regresión no paramétrica son una respuesta a esta limitación, posibilitando que sean los datos los que propongan al analista la forma funcional de las variables predictoras y no al contrario. En un reciente artículo sobre la transferencia de las técnicas estadísticas desde la literatura propiamente estadística a las publicaciones sanitarias, Altman[20] aventura que los modelos aditivos generalizados serán técnicas de uso cada vez más frecuente en la investigación en medicina y salud pública.

El procedimiento LOWESS también es de gran utilidad en los análisis de regresión clásicos; en efecto, una vez ajustado un modelo, siempre es fundamental la realización del diagnóstico para evaluar, entre otras cuestiones, la linealidad de las variables predictoras. Parte del diagnóstico de los modelos clásicos se basa en las relaciones derivadas de nubes de puntos entre los residuales y los valores predichos o las variables predictoras del modelo ajustado. Cuando se utilizan bases de datos suficientemente grandes, es difícil evaluar a simple vista la relación derivada de tales nubes de puntos, por lo que un procedimiento de alisamiento puede ser de inestimable ayuda.

Todos los procedimientos estadísticos modernos, y especialmente la estimación de los modelos de regresión no paramétrica, conllevan una ingente cantidad de cálculo; pertenecen al grupo de los que Efron[21] denomina métodos de computación intensiva. Actualmente existen medios informáticos apropiados en versiones para ordenador personal que sin duda van a posibilitar su utilización. Aunque la oferta es variada, S-PLUS[22] es el más popular, posiblemente porque es una gran herramienta. El hecho de ser programable permite la construcción de *macros* para realizar análisis específicos. También existe un buzón en INTERNET de donde se pueden obtener (también se pueden mandar) de forma gratuita varias de estas *macros.* GAIM es un programa realizado por Hastie y Tibshirani de libre disposición que se puede obtener de sus autores; el precio a pagar es su limitación como paquete estadístico, pues es un programa específico para este tipo de análisis.

Como ocurre para todo método estadístico, la aplicación de estas técnicas deber ser cuidadosa, pero su correcta utilización redundará, sin duda, en un mejor conocimiento de nuestros problemas de salud. En la tabla 1 se presentan, a modo de resumen, diferentes problemas de Salud Pública que han sido estudiados por algunos autores utilizando técnicas de regresión no paramétrica.

Tabla 1. Ejemplos de problemas de salud analizados mediante técnicas de regresión no paramétrica

Estudio	Autores
– Relación entre edad y volumen expiratorio	Segal y cols.[4]
– Relación entre consumo de café y riesgo de infarto de miocardio	Grenland y cols.[7]
– Bajo peso al nacer	Hosmer y Lemeshow[15]
– Muerte neonatal	Herman y Hastie[16]
– Ingresos hospitalarios por neumonía	Schwartz[19]
– Diagnóstico en un modelo de supervivencia para el carcinoma	Aalen[23]
– Relación entre peso de la madre y bajo peso al nader del hijo	Bowman y Young[24]
– Evolución del número de células CD4 en pacientes con HIV	Galai y cols.[25]

Agradecimientos

Los autores agradecen las revisiones críticas hechas por J.M. Antó, E. Perea y M.V. Zunzunegui a anteriores versiones de este manuscrito.

Bibliografía

1. Birkes D, Dodge Y. Alternative methods of regression. New York: Wiley, 1993.
2. Hastie TJ, Tibshirani RJ. Generalized additive models (with discussion). Statist Sci 1986;1:297-318.
3. Friedman JH, Stuetzle W. Smoothing of scatterplots. Techinal Report Orion 003, Dept. of Statitstics. Stanford University, CA. 1982.
4. Segal MR, Weiss ST, Speizer FE, Tager IB. Smoothing methods for epidemiologic analysis. Statist Med 1988;7:601-11.
5. Hardle W. Smoothing techniques with implementation in S. New York: Wiley, 1990.
6. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Statist Ass 1979;74:828-36.
7. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol 1994;140:290-6.
8. Bush NE, Haberman M, Donaldson G, Sullivan KM. Quality of life of 125 adults surviving 6-18 years after bone marrow transplantation. Soc Sci Med 1995;40:479-90.
9. Silverman BW. Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). J R Statist Soc B 1985;47:1-53.
10. Harrell FE Jr, Lee KL, Pollock BG. Regression models in clinical studies: determining relationship between predictors and response. J Natl Cancer Inst 1988;80:1198-202.
11. Durrleman S, Simon R. Flexible regression models with cubic splines. Statist Med 1989;8:551-61.
12. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl Statist 1994;43:429-67.
13. McCullagh P, Nelder JA. Generalized linear models, 2nd ed. London: Chapman and Hall, 1989.
14. Hastie TJ, Tibshirani RJ. Generalized additive models. London: Chapman and Hall, 1990.
15. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley, 1989.
16. Herman AA, Hastie TJ. An analysis of gestacional age, neonatal size and neonatal death using non-parametric logistic regression. J Clin Epidemiol 1990;43:1179-90.
17. Hastie TJ, Tibshirani RJ. Non-parametric logistic and proportional-odds regression. Appl Statist 1987;36:260-76.
18. Hastie TJ, Tibshirani RJ. Exploring the nature of covariates effects in the proportional hazards model. Biometrics 1990;46:1005-6.
19. Schwartz J. Air pollution and hospital admissions for the elderly in Birmingham, Alabama. Am J Epidemiol 1994; 139:589-98.
20. Altman DG, Gooman SN. Transfer of technology from statistical journals to the biomedical literature. JAMA 1994;272:129-32.
21. Efron B. Computer-intensive methods in statistical regression. SIAM Review 1988;30:421-49.
22. S-PLUS. Statistical Sciences Inc. Seattle WA, USA.
23. Aalen OO. Further results on the non-parametric linear regression model in survival analysis. Statistics in Medicine 1993;12:1569-88.
24. Bowman A, Young S. Graphical comparison of nonparametric curves. Appl Statist 1996;45:83-98.
25. Galai N, Muñoz A, Chen K, Carey VJ, Chmiel J, Zhou SY. Tracking of markers and onset of disease among HIV-1 seroconverters. Statistics in Medicine 1993;12:2133-45.