Selección algorítmica de modelos en las aplicaciones biomédicas de la regresión múltiple

Luis Carlos Silva Ayçaguer e Isabel María Barroso Utra*

Investigador titular. Vicerrectoría de Investigación y Posgrado. Instituto Superior de Ciencias Médicas de La Habana.
**Especialista en Bioestadística. Instituto Nacional, Epidemiología y Microbiología. La Habana. Cuba.*

En este artículo nos proponemos valorar la selección algorítmica de submodelos en el contexto de la regresión múltiple, tanto desde una perspectiva teórica como atendiendo al empleo actual que se hace de tal procedimiento. Mediante recursos de simulación se pone de manifiesto la inconsistencia de la regresión paso a paso (RPP), en el sentido de que produce resultados discrepantes tanto cuando se emplean diferentes alternativas dentro de este método como cuando se trabaja con distintas muestras procedentes de la misma población. Tal constatación subraya la incapacidad de este método para identificar factores causales o de riesgo. El examen de la producción científica de dos importantes revistas biomédicas arroja, por una parte, que la RPP casi nunca se usa en la investigación de mayor impacto, así como que sólo raramente se aplica para la única circunstancia en que es pertinente generar funciones pronósticas. Ambos resultados se complementan con varias consideraciones teórico-conceptuales para convencernos de la improcedencia de emplear recursos algorítmicos para subseleccionar modelos explicativos de la realidad en estudio.

La regresión múltiple en sus diversas modalidades (básicamente regresión lineal múltiple, logística, proporcional de Cox y de Poisson) está entre las técnicas estadísticas más usadas en epidemiología. Un examen de la bibliografía especializada lo pone de manifiesto. Por ejemplo, de los 1.178 artículos que abordan problemas prácticos en la *American Journal of Epidemiology* (AJE) entre 1994 y 1998, 855 (casi un 73%) utilizaron al menos una vez el análisis de regresión multivariante[1].

El empleo concreto que se haga de este recurso, como es obvio, puede variar en función de los objetivos del estudio, de la naturaleza de las variables involucradas y de la función matemática supuestamente capaz de describir la relación que las vincula. Las posibles finalidades con que se emplea esta técnica son tres: descriptiva, explicativa y predictiva[2]. Estamos en el primer caso cuando simplemente se quiere explorar la forma en que una variable depende funcionalmente de otras. La segunda concierne al propósito de ayudar a «explicar» o a «entender mejor» los mecanismos o leyes que gobiernan ciertos procesos. Por ejemplo, es posible usarla para esclarecer la relación que vincula la probabilidad de que un niño presente bajo peso al nacer con factores maternos tales como la ganancia de peso durante el embarazo, el número de abortos anteriores y el hábito de fumar[3]. En este contexto tiene particular relevancia la potencialidad de dichos modelos para evaluar los efectos de una variable después de haberse «controlado» el de un conjunto de «factores de confusión». La tercera posibilidad es la de emplearla con propósito predictivo, para vaticinar el comportamiento de cierto fenómeno (variable de respuesta) en función de un conjunto de variables predictivas[4].

Entre los recursos asociados a la regresión múltiple destacan los *métodos de selección algorítmica del modelo*, concebidos para identificar aquellas variables que habrán de integrar la función que a la postre será empleada como modelo resumen del proceso bajo estudio. La lógica subyacente de tal recurso consiste en conservar las variables independientes que contienen información relevante y, a la vez, prescindir de aquellas que resulten redundantes respecto de las que quedaron en el modelo. Estos procedimientos son de índole exclusivamente estadística; discurren según algoritmos programables en los que, una vez elegido el conjunto inicial de variables, no intervienen los juicios teóricos de los investigadores.

Las variables que se retengan estarán, obviamente, entre las que originalmente se hayan elegido para ser incluidas en el estudio. Tal elección inicial siempre padece de un cierto grado de subjetividad, inevitable en toda investigación no experimental, pero a la vez se beneficia de la racionalidad que caracteriza a ese proceso.

Se han ideado varias alternativas para seleccionar un modelo final a partir de este punto. La más conocida es la llamada *regresión paso a paso* –RPP– (*stepwise method*), incorporando variables al modelo (*forward selection*) e ir eliminando variables de él (*backward elimination*). Virtualmente todos los grandes paquetes informáticos para el tratamiento estadístico de datos (tales como SPSS, SAS, BMDP o MINITAB) brindan la posibilidad de aplicar al menos estas dos opciones.

En el método «hacia delante» se comienza calculando los coeficientes de correlación lineal entre cada una de las diversas variables independientes y la dependiente. Luego se identifica aquella que produzca el mayor de estos coeficientes; si a la F correspondiente se le asocia una valor p menor que cierto α_1 prefijado (casi siempre igual a 0,05), esa variable se incluye en el modelo. Si la primera queda incluida (en otro caso, se concluye el proceso sin inclusión alguna) se busca la variable que tenga la mayor correlación parcial con la independiente de entre aquellas que no hayan sido incluidas hasta entonces, y se valora si cumple el criterio de inclusión. El proceso se detiene cuando ninguna de las no incorporadas produzca una F cuyo valor de p sea inferior a α_1.

En el método «hacia atrás» se introducen todas las variables en la ecuación y luego se va considerando la posibilidad de eliminarlas. Se identifica la variable cuya correlación parcial con la variable de respuesta sea la menor; si la F es suficientemente pequeña (probabilidad asociada mayor que cierto α_2, que salvo excepcionales es igual a 0,1), se elimina esa variable y se procede a hacer la misma valoración con la que menor coeficiente de correlación parcial tenga de entre las que aún se conservan. El proceso concluye cuando ya no haya variables que puedan ser eliminadas.

En la práctica no es inusual que se ajuste un modelo de regresión múltiple y de inmediato se aplique un procedi-

Correspondencia: Dr. L.C. Silva Ayçaguer.
Vicerrectoría de Investigación y Posgrado.
Instituto Superior de Ciencias Médicas de La Habana.
Calle G y 25, sexto piso, Plaza, Ciudad de La Habana.
Correo electrónico: lcsilva@intomed.sld.cv

Recibido el 30-8-2000; aceptado para su pulicación el 22-1-2001

Med Clin (Barc) 2001; 116: 741-745

miento algorítmico para determinar qué variables han de «quedarse» en calidad de factores detectados como verdaderamente influyentes y cuáles habrán de desecharse[5].
El empleo de estos recursos con fines explicativos es, como mínimo, muy discutible. Sus resultados suelen ser interpretados como sigue: las variables que se «quedan» dentro del modelo final son las causantes (y quizá las principales causantes) de las modificaciones que experimenta la variable dependiente; las que no permanecen, o bien no influyen causalmente en el proceso, o su influencia no es apreciable. En efecto, muchos investigadores utilizan la selección algorítmica de modelos con la aspiración de obtener de manera automática conclusiones explicativas sobre el proceso causal que estudian.
La bibliografía especializada[2-4,6] relacionada con la estrategia de selección de variables no aporta argumento convincente alguno que permita considerar que ir «hacia delante» sea mejor que ir «hacia atrás», o viceversa. Tal circunstancia constituye un primer indicio de la posible improcedencia de confiar a un algoritmo como la RPP la tarea de explicar la realidad, puesto que es legítimo sospechar que las variables que conforman el modelo final podrían ser las mismas para ambos procedimientos. Tal desempeño inconsciente de la RPP se ha reflejado en la bibliografía; por ejemplo, McGee et al[7] exponen detalladamente un ejemplo basado en la regresión logística en que cada uno de los tres procedimientos de selección algorítmica produce resultados finales drásticamente diferentes entre sí.
En nuestra opinión, la esperanza de que el empleo de estos procedimientos contribuya a «entender» o «explicar» la realidad es, en el mejor de los casos, estéril o quimérica y, no con baja probabilidad, contraproducente. El objetivo del presente artículo es fundamentar tal convicción tanto teórica como prácticamente.

Material y método

Se abordaron dos aspectos relacionados con la práctica: por una parte, se examinó el «funcionamiento» de las técnicas algorítmicas por medio de datos hipotéticos, y por otra, se indagó en la bibliografía científica contemporánea para valorar cuantitativa y cualitativamente el empleo de dichas técnicas.

Estudio con datos simulados

La idea básica consistió en generar bases de datos compatibles con cierta estructura teóricamente prefijada y evaluar la posible inconsistencia de distintos procedimientos algorítmicos a los efectos de delimitar «modelos finales» de regresión. Puesto que para los fines ilustrativos de este estudio cualquiera de las modalidades de la regresión múltiple son, en principio, esencialmente equivalentes, seleccionamos la más simple y familiar de ellas: la regresión lineal.
Cabe aclarar que el recurso de simulación no se empleó en el sentido de los estudios tipo Montecarlo –tal y como se exponen en los libros clásicos[8]– con la finalidad de aquilatar el grado de eficiencia con que se desempeña un procedimiento inferencial específico, sino que se empleó con la intención exclusiva de conformar unos pocos juegos de datos que pusieran en evidencia su improcedencia. La lógica con que se construyeron las ilustraciones tiene en cuenta que, por muchas confirmaciones que se consigan de que una afirmación es cierta, ello no basta para darla por demostrada, a la vez que un solo contraejemplo, en cambio, es suficiente para descartarla como correcta. Quiere esto decir que si hallamos bases de datos hipotéticos que, tras el empleo de la selección algorítmica de modelos, conduzcan a conclusiones explicativas contradictorias o absurdas, entonces tal recurso quedaría en entredicho.
La idea concreta en este caso fue la de simular diferentes juegos de datos, todos compatibles con el mismo modelo estructural, y someterlos a distintas variantes de selección algorítmica de variables. Si los resultados finales (los modelos resultantes) para diferentes variantes de selección de modelos aplicadas a un mismo juego de datos, o cuando se aplica la misma variante de selección algorítmica a diferentes muestras, son muy diferentes entre sí, entonces tendríamos evidencias tangibles de la inconsistencia del método.
Para construir las muestras simuladas, se trabajó con el modelo siguiente:

$$Y = \alpha + \sum_{i=1}^{k} \beta_i X_i \quad [1]$$

Fijados α, β_1, ..., β_k, la siguiente tarea estriba en obtener, mediante simulación, una matriz de datos consistente con dicha ecuación. Esto es, generar un conjunto de n vectores «empíricos», cada uno del tipo (Y_j, X_{1j}, X_{2j}, ..., X_{kj}), que constituyan realizaciones aleatorias del vector (Y, X_1, X_2, ..., X_k), donde $j = 1, \cdots, n$. Para ello es preciso completar cuatro pasos: imponer una distribución al subvector (X_1, X_2, ..., X_k), generar n expresiones aleatorias del mismo de acuerdo con tal distribución, computar [1] para cada una de ellas y, finalmente, añadir un error aleatorio ε_j al número resultante en cada caso para obtener los Y_j.
Se decidió operar con k = 6 variables continuas, distribuidas normalmente con medias 5, 10, 15, 20, 25, 30 y desviaciones estándar iguales a 0,5, 1,0, 1,5, 2,0, 2,5 y 3,0, respectivamente (esto equivale a imponer que los coeficientes de variación fueran constantes: $\sigma_i/\mu_i = 0,1$). Se prefijó asimismo una matriz de correlaciones entre las X_i con la siguiente estructura:

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1					
X_2	0,9	1				
X_3	0,9	0,8	1			
X_4	0	0	0	1		
X_5	0	0	0	0,9	1	
X_6	0	0	0	0,9	0,8	1

Como se aprecia, se conformaron dos subgrupos, cada uno integrado por tres variables: X_1, X_2 y X_3 con una alta correlación mutua y, a la vez, estructuralmente incorrelacionadas con las variables del segundo grupo, conformado éste por X_4, X_5 y X_6, las cuales también están altamente correlacionadas entre sí.
Para contar con algún modelo general que permitiera el análisis, se eligieron 7 coeficientes de regresión. La función seleccionada, en definitiva, fue*:

$$Y = 50 - X_1 - 0,5\ X_2 - 0,333\ X - 0,25\ X_4 - 0,2\ X_5 - 0,166\ X_6 \quad [2]$$

En resumen, sobre estas bases, para cada vector (X_{1j}, X_{2j}, ⋯, X_{6j}) se obtuvo $Y_j = \alpha + \sum_{i=1}^{6} \beta_i X_{ij} + \varepsilon_j$, para lo cual se decidió que ε_j fuera una realización de la distribución normal con media $\mu = 0$ y desviación estándar $\sigma = 3,5$ (j = 1, ⋯, n).
Finalmente, decidimos generar n = 500 casos, un tamaño muestral compatible con el de una aplicación típica en la práctica. Debe advertirse que una muestra muy voluminosa (p. ej., de 2.000) haría muy potentes las pruebas de significación involucradas en el proceso de selección de variables (conduciendo quizá a que todas las variables fueran incluidas siempre). En este sentido es fundamental que recordemos que para rechazar una hipótesis nula basta tomar una muestra suficientemente grande (salvo en el caso excepcional en que el coeficiente de regresión sea exactamente igual a cero, lo cual, por definición, no se produce aquí), realidad sintetizada por Savage[9] en su ya célebre afirmación: «Con mucha frecuencia se sabe de antemano que las hipótesis de nulidad son falsas incluso antes de recoger los datos; en ese caso el rechazo o la aceptación simplemente es un reflejo del tamaño de la muestra».
Así se obtiene una matriz de datos coherente con el modelo [2]. Para conferir más consistencia al examen del problema, se repitió cinco veces ese proceso. Se obtuvieron, por tanto, 5 juegos de 500 datos cada uno, a todos los cuales se les ajustó la regresión lineal múltiple. A continuación se aplicó la selección algorítmica de modelos a cada uno de ellos según las dos variantes previstas: *backward selection* (selección hacia atrás), *forward selection* (selección hacia delante), siempre con niveles de significación de entrada y de salida de las variables iguales a $\alpha_1 = 0,05$ y $\alpha_2 = 0,10$, respectivamente, que son los asumidos por defecto en el programa empleado (SPSS para Windows).
En principio, los 10 resultados deberían ser esencialmente iguales. Si los resultados hacia atrás y hacia delante discrepan de forma notable para un mismo juego de datos, ello indicará que la selección algorítmica carece de una consistencia que resultará vital cuando se trate de distinguir entre las relaciones causales y las que se producen por otras razones (p. ej., debido a los llamados factores de confusión, que se caracterizan por tener un grado alto de asociación tanto con el agente causal como con el efecto). Por otra parte, si la misma estrategia de selección algorítmica produjera resultados muy divergentes para distintas representaciones muestrales del mismo modelo general, esto querrá decir que el procedimiento es altamente sensible a variaciones muestrales y, por tanto, poco robusto ante tal contingencia y obviamente inevitable en la práctica.
Procede subrayar que se ha trabajado con una larga serie de decisiones (tipo de regresión, número de variables, tamaño muestral, estructura de correlación, valores de los coeficientes, etc.) adoptada del modo expuesto por razones de mera conveniencia operacional. Nótese que se ha procedido como si se quisiera evaluar la fiabilidad de un instrumento de medición (p. ej., de un cuentakilómetros): para cuantificarla podría ser recomendable elegir varios tramos de carretera y recorrer cada uno de ellos varias docenas de veces con la finalidad de estimar la consistencia. Sin embargo, para probar que dicho instrumento es inútil bastaría corroborar que al realizar 5 viajes de ida y vuelta entre dos puntos prefijados concretos, se obtienen 10 mediciones radicalmente diferentes entre sí.

Examen de la bibliografía

El marco de la discusión resultaba propicio para acudir a algún marco referencial que permitiera una valoración más integral del tema. Se decidió realizar un examen bibliográfico que permitiera aquilatar el grado y el modo en que se emplean las técnicas que nos ocupan en el ámbito epidemiológico.

*Los coeficientes se tomaron de manera que $E(\beta_i X_i) \approx 5$ y, por tanto, $E(Y) \approx 20$, aunque no existió ninguna razón especial para ello.

Una de las fuentes elegidas fue la *American Journal of Epidemiology* (AJE), publicación de primera línea mundial y de notable impacto en el área de la epidemiología. Lo que allí se publica es, a nuestro juicio, un reflejo de lo que pudiéramos considerar un paradigma o estándar metodológico adecuado, pues aquellos artículos que consiguen ser admitidos en revistas como ésta se han sometido a un examen especialmente exigente y riguroso. Es cierto que no todos los trabajos acogidos en este tipo de revistas son necesariamente impecables, ni son obligadamente mediocres los que terminan en una de segundo o tercer orden; pero a nuestro juicio no será fácil hallar patrones de comparación más adecuados. Para esta valoración se eligió el período 1994-1998, ambos años inclusive. Por otra parte, se decidió sondear la situación prevalente en una revista también prestigiosa, aunque de mucho menor impacto y acaso más expuesta al empleo acrítico de recursos estadísticos cuestionables; se eligió la revista española MEDICINA CLÍNICA (MC).

Se revisaron en su totalidad los artículos donde se usó alguna forma de regresión múltiple. El sondeo de la revista MEDICINA CLÍNICA se realizó empleando una base de datos contenida en un CD-ROM distribuido por Doyma y *Pharmacia and Upjohn*, que contiene todos los artículos allí publicados entre 1992 y 1998, período similar al elegido para AJE. La expectativa era hallar muy pocos trabajos en AJE que emplearan RPP y, a la vez, un uso apreciablemente más intenso en MC.

Por último, para el examen cualitativo de la situación se realizó una valoración crítica de todos y cada uno de los artículos que emplearon la selección de modelo paso a paso en ambas revistas.

Resultados

Análisis por simulación

Tras aplicar un programa de simulación de vectores aleatorios creados *ad hoc* para cumplimentar la tarea planificada, se obtuvieron las 5 bases de datos previstas, que pueden solicitarse por correo electrónico a los autores. Será fácil comprobar para quienes lo deseen que las X_1 siguen la distribución multinormal estipulada y que los datos se ajustan adecuadamente a la combinación lineal [2].

Los resultados obtenidos al aplicar estos dos procedimientos de selección a cada una de las 5 bases de datos se recogen en la tabla 1.

Tal como esperábamos, cada base de datos produjo su propio «desenlace» y el panorama general es simplemente caótico: por una parte, para ninguna de las 5 muestras los dos procedimientos algorítmicos produjeron el mismo modelo final, y por otra, cada uno de los dos procedimientos dio lugar a 5 modelos finales distintos cuando se emplearon diferentes representaciones muestrales. Por añadidura, cada una de las 6 variables fue elegida en al menos una de las 10 ocasiones, a la vez que ninguna de ellas resultó incluida en la totalidad de las experiencias.

Empleo de métodos algorítmicos de selección en la producción científica contemporánea

La regresión paso a paso tiene una presencia totalmente marginal como recurso de análisis entre los trabajos que emplearon regresión múltiple en AJE. La situación detallada se recoge en la tabla 2.

El resultado que arrojó el mismo examen para MEDICINA CLÍNICA (tabla 3) demuestra un patrón de empleo no muy acusado pero notablemente superior al de *American Journal of Epidemiology*: en MC esta técnica se usa 11 veces más que en AJE. Especialmente notable es la diferencia registrada para la regresión lineal múltiple: mientras que en AJE ni uno solo de los 142 artículos complementó su análisis de regresión con RPP, en MC se empleó tal aderezo en uno de cada 10.

Desde el punto de vista cualitativo, valoramos detalladamente los 9 artículos de AJE y los 38 de MC que emplean RPP. A nuestro juicio, sólo dos de los 9 artículos que utilizan tal recurso en AJE lo hacen de manera no objetable[10,11], a la vez que hay otros dos[12,13] en que al menos se da una explicación explícita y transparente de por qué emplean la RPP, aunque no sea con la finalidad exclusiva de construir un modelo predictivo. En los restantes 5, dicho empleo es erróneo.

TABLA 1

Variables que quedaron incluidas al aplicar la regresión paso a paso hacia atrás y hacia delante a cada uno de los 5 juegos de datos

	Hacia atrás						Hacia delante					
	X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6
Muestra 1		*	*	*			*			*		
Muestra 2			*		*	*			*	*		
Muestra 3		*	*			*	*	*				*
Muestra 4		*	*	*		*	*					*
Muestra 5		*	*			*			*	*		

TABLA 2

Tasas de empleo de la regresión paso a paso según modalidades de regresión en *American Journal of Epidemiology*, 1994-1998

Tipos de análisis de regresión múltiple	Artículos que utilizan el análisis de regresión	Artículos que utilizan la selección paso a paso	Porcentaje
Logística	477	6	1,3
Cox	188	2	1,1
Lineal	142	0	0,0
Poisson	48	1	2,1
Total de artículos	355	9	1,0

TABLA 3

Tasas de empleo de la regresión paso a paso según modalidades de regresión múltiple en MEDICINA CLÍNICA, 1992-1998

Tipos de análisis de regresión múltiple	Artículos que utilizan el análisis de regresión	Artículo que utilizan la selección paso a paso	Porcentaje
Logística	149	24	16,1
Cox	49	2	4,1
Lineal	117	11	9,4
Poisson	21	1	4,8
Total de artículos	336	38	11,3

En el caso de MC la inmensa mayoría de los 38 trabajos que utilizan la RPP lo hacen para identificar variables relevantes en la determinación causal de los fenómenos: sólo se identifican 4 que utilizan el procedimiento algorítmico con propósitos verdaderamente predictivos[14-17]. En general, tanto en una revista como en la otra se registran explicaciones muy confusas acerca de los propósitos con que se emplea el método algorítmico. A ello contribuye el hecho de que dentro del mismo texto, antecediendo a la expresión que alude a la enfermedad o problema que se estudia, se utilicen las más diversas expresiones para calificar o denominar a las variables que son objeto de selección algorítmica («predictores», «variables asociadas con...», «factores que tienen un efecto sobre...», «factores de riesgo», «factores predictivos del riesgo», «mediadores», «variables influyentes en...», «variables explicativas de...», «variables estadísticamente responsables de...», «variables relacionadas con...», «factores más decisivos en...» y «factores pronósticos de...» son algunas de las variantes empleadas). Este fenómeno se presenta de manera algo más acusada en MC que en AJE. Lo cierto es que en la mayoría de los casos la RPP se usa en definitiva para identificar factores de riesgo o descubrir asociaciones que se interpretan como indicación de causalidad.

Discusión

Las discrepancias resultantes al aplicar las dos variantes algorítmicas de la RPP revelan la notable inconsistencia del méto-

do y ponen de manifiesto que no podrá ser capaz de identificar las variables que *expliquen* el proceso modelado. Es muy probable que dos investigadores, empleando el mismo recurso general pero diferentes alternativas específicas igualmente aceptables a la hora de concretarlo (uno hacia atrás y otro hacia delante), lleguen a conclusiones radicalmente discrepantes al identificar cuáles son las variables causantes de las variaciones que experimenta la variable dependiente (Y).

Por otra parte, al aplicar un mismo procedimiento algorítmico a diferentes representaciones muestrales de la «realidad», los modelos finales obtenidos también fueron estructuralmente diferentes. Debe repararse en que dichos juegos de datos son estadísticamente equivalentes; las diferencias que exhiben son debidas a las variaciones aleatorias propias del muestreo (tal como ocurriría a diversos investigadores que estudiaran la misma población por medio de respectivas muestras). Vale decir que 5 investigadores independientes que estén encarando la misma realidad y empleando el mismo método, y con el mismo tamaño muestral, llegarían a conclusiones diferentes entre sí sólo por el hecho de trabajar con muestras diferentes.

Para interpretar más claramente estos resultados, consideremos el problema desde una perspectiva práctica. Supongamos que un investigador estuviera interesado en conocer cuáles son los factores que verdaderamente influyen en el padecimiento de enfisema pulmonar antes de los 60 años de edad. Para ello podría comenzar seleccionando un conjunto de variables presuntamente explicativas del proceso que se estudia, como la edad, los antecedentes patológicos familiares, el hábito de fumar cigarrillos, la pigmentación de los dedos (pulgar, índice y medio) de la mano, la polución ambiental en su medio laboral, etc. Podría entonces realizar un estudio prospectivo, aplicar la regresión logística y, finalmente, llevar a cabo una selección mediante RPP con la esperanza de descubrir cuáles de estas variables tienen mayor peso causal sobre la aparición de enfisema (o son factores que entrañan mayor riesgo de desarrollar la dolencia).

Como ponen de manifiesto los resultados, no sería para nada inverosímil que tal acción tuviera como posible resultado que el grado de pigmentación en los dedos quedara incluido en el modelo y que no ocurriera lo mismo con la condición de fumador, debido a la madeja de correlaciones que tienen estas dos variables con las restantes del modelo y a la asociación que tienen entre sí. Si no fuera por la certeza que existe hoy de que el hábito de fumar es una práctica que favorece el enfisema pulmonar y de que la pigmentación de los dedos (inducida por el hábito) es una variable de confusión para esta relación causal, el investigador sería conducido por sus propias reglas de análisis a aseverar que la pigmentación de los dedos, en caso de que quedara incluida en el modelo final, es un factor de rieso para esta entidad. Tal conclusión, lejos de iluminar el camino hacia el conocimiento de las verdaderas relaciones causales, lo ensombrecería o, más bien, lo obstaculizaría. Conviene no olvidar que «los número no saben de dónde vienen»[18].

Debe advertirse que existe una notable ambivalencia cuando se habla de «variables predictivas» en situaciones que no son verdaderamente de predicción. «Factor de riesgo» y «factor predictivo» no son sinónimos: el hecho de que esté o no presente un factor de riesgo (y el grado en que gravita en una función) puede ocaionalmente ser útil para la predicción, pero una variable puede hacer importante aportación a los efectos de predecir, aunque en sí misma no sea un factor de riesgo.

Volviendo al ejemplo anterior, si el modelo de regresión se aplicara para estimar la probabilidad de que un sujeto con cierto perfil desarrolle en el futuro un enfisema pulmonar (p. ej., para emprender una especial acción preventiva sobre quienes tengan tal perfil), entonces la pigmentación de los dedos podría estar con todo derecho en él, ya que en este caso el enfoque ha de ser pragmático: si se consiguen buenas predicciones, poco importan los medios. Es en este contexto irrelevante si una variable dada ha quedado incluida porque desempeña un papel causal, o por ser un mero reflejo de otra que no aparece pero que sí pudiera tener tal condición; lo que importa es construir el instrumento predictor con el menor número posible de variables, principio de parsimonia que reducirá los esfuerzos que ha de realizar el investigador, tanto en la recogida de la información como en el manejo ulterior de la ecuación. Pero si se quiere *entender* el mecanismo de producción del enfisema, no por calificar de «predictivas» a las variables independientes la situación pasa a ser *de predicción*; de manera que resulta absurdo actuar como si el término diera amparo a un procedimiento esencialmente inconducente (la RPP) para lo que realmente se está haciendo, que es identificar factores causales o de riesgo.

Otra trampa semántica que contribuye al empleo equivocado de la RPP concierne al hecho de que dos variables estén asociadas o no. Tras aplicar este procedimiento, muchos investigadores «concluyen» que las variables que quedaron en el modelo *están asociadas* con el fenómeno que se estudia (típicamente una enfermedad), con lo que eluden el compromiso de pronunciarse acerca de si tales variables son o no *causantes* del fenómeno en cuestión. No tiene sentido convertir la constatación de que dos variables están asociadas en una *conclusión*, porque esta última debe ser el resultado de un proceso intelectual cualitativamente superior a la mera cuantificación fenomenológica que la primera representa. La constatación la puede hacer el SPSS; a la conclusión sólo puede llegar un ser humano. Nótese además que la cuantificación de una asociación nunca tiene interés en sí mismo y carece de un sentido claro salvo que se inserte en el contexto de una conjetura causal[19]. A nadie se le ocurriría investigar, por ejemplo, la asociación entre la condición de ser hipertenso y el color de la vivienda del paciente; es decir, siempre que se mide una asociación es porque hay una sospecha racional, como mínimo subconsciente, de que tal asociación pudiera brindar una prueba o, al menos, un indicio de una relación causal.

Lo curioso es que, a la vez que muchos autores están avizorados acerca de que no deben confundir asociación con causalidad en el marco univariable, parecen olvidarlo cuando quedan encandilados por los métodos multivariables, y que no comprendan que, al aplicar estos algoritmos mecánicamente, están incurriendo solapadamente en el viejo sofisma. Variables que pudieran tener responsabilidad «directamente causal» pueden resultar eliminadas al ser suplidas por una o más variables que no tengan influencia real alguna, pero que se vinculen con ella, y en la medida que el asunto se dirime en la caja negra de la RPP, nada podemos hacer para evitarlo.

Por otra parte, cabe recordar que estos criterios están asentados sobre las pruebas de significación; por tanto, su pertinencia está sujeta a todas las suspicacias que ellas despiertan[5]; en particular, a la mayor de todas: cuando la muestra es suficientemente grande, cualquier variable quedaría incluida en el modelo, con independencia de que su sustantividad clínica o biológica sea nimia o no.

Como ha dicho Guttman[20], «el uso de la regresión paso a paso es en la actualidad una confesión de ignorancia teórica sobre la estructura de la matriz de correlaciones». Cuando la regresión múltiple se usa para describir los patrones de causalidad según los cuales ciertas variables actúan so-

bre otra, la regresión paso a paso equivale a cubrir esa ignorancia con un algoritmo que piense por el investigador. No en balde el *stepwise regression* fue rebautizado irónicamente[21] como *unwise regression* (juego de palabras intraducible que aproveha que el vocablo *wise* denota en inglés la manera o el modo de hacer algo, pero también significa «sabio», de modo que *unwise regression* vendría a ser algo así como regresión tonta o irracional).

La aplicación de la RPP dimana en muchos casos de la creencia de que complejos modelos estadísticos multivariables podrían desentrañar las posibles causas de aquellas enfermedades muy complicadas y dependientes de un gran número de variables mutuamente correlacionadas, como ocurre con las dolencias coronarias y los tumores malignos. De esta forma se ignora que la dificultad no radica en que se estén aplicando métodos estadísticos insuficientemente alambicados, sino en la falta de teorización.

Tal ingenuidad ha sido advertida[22] aunque, a nuestro juicio, quizá no con la suficiente intensidad. Se ha destacado la necesidad de incluir sólo variables cuyo sentido epidemiológico o clínico esté claro, pero eso no resuelve el problema cardinal: si aspiramos a que la RPP nos conduzca a conseguir conocimientos que no teníamos sobre la preeminencia causal de unas variables sobre otras, necesariamente tendremos que incluir variables iniciales cuyo papel ignoramos; y viceversa: si conocemos cabalmente ese papel, entonces no puede decirnos nada novedoso, y la selección algorítmica del modelo se convierte en una finalidad, no en un medio. No es posible escapar de este laberinto. Los esfuerzos por resolver el problema a través de poderosos programas informáticos recuerdan a los constructores de máquinas de movimiento perpetuo: ignorantes de la ley de conservación de la energía y creyendo que sus fracasos se debían a que el diseño del aparato no era suficientemente ingenioso, procedían a desgastarse en la confección de nuevos y más sofisticados dispositivos[23].

En este contexto, cabe hacer una apelación al sentido común: si el empleo mecánico de recursos estadísticos multivariables pudiera ayudar a esclarecer las complejas relaciones causales que expliquen por qué unos individuos enferman y otros no, entonces con las enormes bases de datos hoy disponibles, las poderosísimas y veloces computadoras actuales y los potentes programas informáticos a los que todos tenemos acceso, la etiología del cáncer de mama, por poner un ejemplo, no sería el misterio que es hoy para la ciencia y que obliga a la generación apremiante de nuevos enfoques[24], muy alejados de la acomodaticia esperanza de que el *software* puede suplir nuestra perspicacia y creatividad.

El patrón general de empleo de la RPP fuera del marco estrictamente predictivo en AJE es elocuente: los investigadores de más nivel no consideran por lo general que este recurso pueda contribuir a explicar nada. De hecho, su empleo es completamente marginal: sólo 9 de 789 usuarios de la regresión múltiple acuden a él y la mayoría (5) lo emplea de manera *unwise*. Un grupo bastante numeroso de los trabajos que emplean la regresión en MC recurren a la RPP (38 de 336). La inmensa mayoría (34 de esos 38) no parece consciente de sus limitaciones como fuente explicativa. Se dan incluso casos en que tal empleo mecánico llega al extremo de calificar a una sola variable (la única que quedó en el modelo tras la aplicación de la RPP) como «la única que influye» en la aparición de la dolencia[25,26].

En síntesis, el problema fundamental está en la pretenciosa y a la vez ingenua interpretación que suele hacerse del resultado que arroja la RPP. Su empleo con fines explicativos es absurdo, pues la selección algorítmica de modelos no puede evitar que los resultados se deriven de meras concomitancias estadísticas (de hecho, en eso se basan), ni distinguir entre las asociaciones de índole causal y las debidas a terceros factores involucrados en el proceso. Consecuentemente, si bien los modelos de regresión múltiple pueden ser de extraordinario interés para ayudar a entender los procesos biológicos y sociales, los procedimientos algorítmicos de subselección de variables para conformar un modelo «final» explicativo son, salvo situaciones excepcionales, totalmente improcedentes.

REFERENCIAS BIBLIOGRÁFICAS

1. Guevara A. Un aporte a la valoración crítica de las tecnologías cuantitativas disponibles en el contexto de la investigación epidemiológica actual. Trabajo de especialista de primer grado en Bioestadística. La Habana: Facultad de Salud Pública, 1999.
2. Chatterjee S, Hadi AS, Price B. Regression analysis by example. Nueva York: Willey, 2000.
3. Silva LC. Excursión a la regresión logística en ciencias de la salud. Madrid: Díaz de Santos, 1995.
4. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied regression analysis and multivariable methods (3.ª ed.). EE.UU.: Duxbury Press, 1997.
5. Silva LC. Cultura estadística e investigación en el campo de la salud. Madrid: Díaz de Santos, 1996.
6. Draper N, Smith H. Applied regression analysis (2.ª ed.). Nueva York: Willey and Sons, 1981.
7. McGee DL, Reed D, Yano K. The results of logistic analyses when the variables are highly correlated. Am J Epidemiol 1984; 37: 713-719.
8. Mirham GA. Simulation. Statistical foundations and methodology. Nueva York: Academic Press, 1972.
9. Savage IR. Nonparametric statistics. J Am Statistic Assoc 1957; 52: 332-333.
10. Myers L, Coughlin SS, Webber LS, Srinivasan SR, Berenson GS. Prediction of adult cardiovascular multifactorial risk status from childhood risk factor levels. Am J Epidemiol 1995; 142: 981-924.
11. Saghavi DM, Gilman RH, Lescano-Guevara AG, Checkley W, Cabrera LZ, Cardenas V. Hyperendemic pulmonary tuberculosis in a peruvian shantytown. Am J Epidemiol 1998; 148: 581-593.
12. Hilton JF, Donegan E, Katz MH, Canchola Aj, Fusaro RE, Greenspan D et al. Development of oral lesions in human immunodeficiency virus-infected transfusion recipients and hemophilliacs. Am J Epidemiol 1997; 145: 164-174.
13. Friedman SR, Jose B, Deren S, Des Jarlais DC, NEaigus A. Risk factors for human immunodeficiency virus seroconversion among out-of-treatment drug injectors in high and low seroprevalence cities. Am J Epidemiol 1995; 142: 864-874.
14. Vilalta J, Vaqué J, Olona M, Castaño CH, Guitart JM, Rosselló J et al. Factores predictivos de la mortalidad en los traumatismos craneoencefálicos graves. Med Clin (Barc) 1992; 99: 441-443.
15. Recasens MA, Barenys M, Fernández-Ballart J, Sol R, Blanch S, Salas Salvadó J. Estimación del gasto energético en pacientes con obesidad mórbida. Med Clin (Barc) 1994; 99: 451-455.
16. Blanch S, Recasens MA, Solá R, Salas-Salvadó J. Efecto de una dieta altamente hipocalórica sobre el control de la obesidad mórbida a corto y medio plazo. Med Clin (Barc) 1993; 100: 450-453.
17. López JA, Armengol O, Chavarren J, Dorado C. Una ecuación antropométrica para la determinación del porcentaje de grasa corporal en varones jóvenes de la población canaria. Med Clin (Barc) 1997; 108: 207-213.
18. Lord FM. On the statistical treatment of football numbers. Am Psychol 1953; 8: 750-751.
19. Silva LC, Benavides A. Causalidad e inobservancia de la premisa de precedencia temporal en la investigación biomédica. Methodologica (Bélgica) 1999; 7: 1-11.
20. Gutman L. What is not what in statistics. The Statistician 1977; 26: 81-107.
21. Leamer EE. Sensitivity analysis would help. Am Econ Rev 1985; 75: 308-313.
22. Doménech JM, Sarriá A. Análisis multivariante en ciencias de la salud. Modelos de regresión. Unidad didáctica 10. Barcelona: Signo, 1995.
23. Silva LC. Hacia una cultura epidemiológica revitalizada. Dimensión Humana 1997; 1: 23-33.
24. Evans RG, Morris LB, Marmor TR. Why are some people healthy and others not? The determinants of health of populations. Nueva York: Aldine Gruyter, 1994.
25. Bastida MT, Martínez JA, López P, Ribera L, Expósito M. Infección urinaria bacteriémica en el varón. Estudio comparativo frente a la pielonefritis bacteriémica femenina. Med Clin (Barc) 1997; 109: 321-323.
26. López L, Portero JA, Borrego D, Báez A, San Miguel JF. Alto gasto cardíaco en pacientes con mieloma. Prevalencia y características clínicas. Med Clin (Barc) 1997; 108: 214-216.