Construcción de modelos de regresión multivariantes

Preparado por Luis M. Molinero (Alce Ingeniería) Abril 2002

Se conoce como análisis de regresión multivariante al método estadístico que permite establecer una relación matemática entre un conjunto de variables X_1, X_2 .. X_k (covariantes o factores) y una variable dependiente *Y*. Se utiliza fundamentalmente en estudios en los que no se puede controlar por diseño los valores de las variables independientes, como suele ocurrir en los estudios epidemiológicos y observacionales.

Los objetivos de un modelo de regresión puede ser dos:

- Obtener una ecuación que nos permita "predecir" el valor de Y una vez conocidos los valores de X_1, X_2 .. X_k. Se conocen como **modelos predictivos**.
- Cuantificar la relación entre X_1, X_2 .. X_k y la variable Y con el fin de conocer o explicar mejor los mecanismos de esa relación. Se trata de **modelos explicativos**, muy utilizados cuando se busca encontrar qué variables afectan a los valores de un parámetro fisiológico, o cuáles son los posibles factores de riesgo que pueden influir en la probabilidad de que se desarrolle una patología.

La disponibilidad y facilidad de uso del software que permite la construcción de modelos de regresión nos ha hecho olvidar que se trata de técnicas complejas, que requieren un cierto conocimiento de la metodología estadística subyacente, por lo que nos encontramos con excesiva frecuencia una pobre utilización de las técnicas de regresión y una peor descripción de cómo se emplearon en cada caso concreto, e incluso una ausencia total de esa explicación, y se comunica los resultados como si la propia ecuación de regresión fuera sin más un "artículo de fe" que no necesitara de una cuidadosa validación.

Un problema fundamental que se plantea a la hora de construir un modelo multivariante es qué factores X_1, X_2 .. X_k incluir en la ecuación, de tal manera que estimemos el mejor modelo posible a partir de los datos de nuestro estudio. Para ello lo primero que habría que definir es qué entendemos por "*mejor modelo*". Si buscamos un modelo predictivo será aquél que nos proporcione predicciones más fiables, más acertadas; mientras que si nuestro objetivo es construir un modelo explicativo, buscaremos que las estimaciones de los coeficientes de la ecuación sean precisas, ya que a partir de ellas vamos a efectuar nuestras deducciones. Cumplidos esos objetivos es claro que otra característica deseable de nuestro modelo es que sea lo más sencillo posible.

Variable de confusión

En el área de los modelos explicativos aparece un concepto de gran importancia, el de **variable de confusión**. Se dice que existe "confusión" cuando la relación entre dos variables difiere de forma importante si se considera el efecto de una tercera, alterando por tanto de alguna manera la interpretación de esa relación.

Veamos un ejemplo. Si estamos estudiando mediante una muestra aleatoria una población de diabéticos y analizamos la posible relación entre la PAS y la edad y sexo de los pacientes, obtenemos mediante un modelo de regresión lineal la siguiente ecuación

Término	Coef.	Err.est.	t	p
Constante	116,285	2,8410	40,931	0,0000
EDAD	0,328	0,0432	7,592	0,0000
SEXO	**2,042**	1,0486	1,947	0,0515

donde la variable SEXO se ha codificado como 0 para los hombres y 1 para las mujeres, de tal manera que el cambio medio de la PAS, estimado por esta ecuación, cuando comparamos a los hombres y a las mujeres manteniendo fija la edad, es de aproximadamente de 2 mmHg (p = 0.052).

Sin embargo si controlamos también el índice de masa corporal (IMC) introduciéndolo en la ecuación, obtenemos:

Término	Coef.	Err.est.	t	p
Constante	101,834	4,0727	25,004	0,0000
EDAD	0,321	0,0426	7,531	0,0000
SEXO	**1,387**	1,0428	1,330	0,1835
IMC	0,514	0,1051	4,889	0,000001

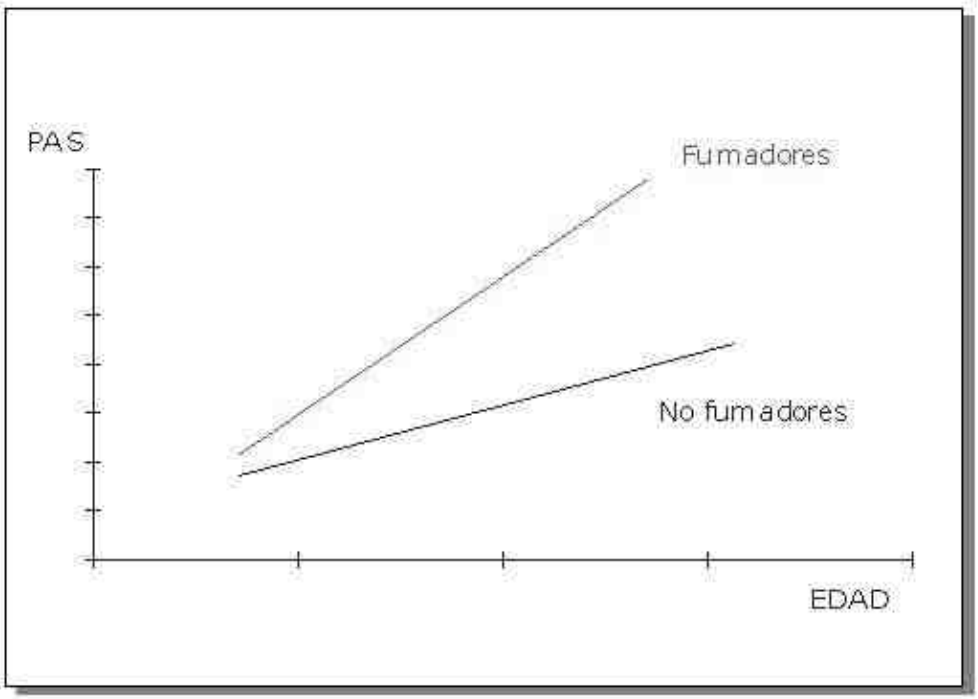
donde comprobamos que al incluir el índice de masa corporal IMC, el coeficiente de regresión de la variable SEXO se ha modificado drásticamente (ha disminuido en más del 30 %), lo que quizás modifica nuestra interpretación de la relación, ya que si se tiene en cuenta el IMC la influencia del sexo no es relevante. En este caso el IMC sería un factor de confusión que deberíamos incluir en la ecuación y ello aunque, al contrario de lo que ahora ocurre, su coeficiente no fuera significativo.

En la práctica habitual vemos que el criterio que se utiliza −incorrectamente− para incluir o no en la ecuación una posible variable de confusión se basa en comprobar si el coeficiente correspondiente es significativamente diferente de 0, para lo cual sólo se mira el valor de la probabilidad asociado a ese contraste. Sin embargo no es esa la única cuestión, sino si su introducción en la ecuación modifica apreciablemente o no la relación entre la variable dependiente y el otro factor o factores estudiados. Se trata pues de utilizar un enfoque clínico o fisiológico, ya que hay que determinar desde ese punto de vista qué consideramos como cambio apreciable en el coeficiente de la ecuación de regresión.

Interacción

Un segundo concepto importante es el de interacción. Decimos que existe **interacción** en la relación entre dos variables cuando los valores de una tercera afectan a esa relación, magnificándola o disminuyéndola, o más raramente ambas cosas dependiendo del nivel de la tercera variable. Es decir que la magnitud de la relación es diferente según los niveles de esa tercera variable.

Así, por ejemplo, podríamos encontrar que la media de la PAS aumenta con la edad, pero que ese aumento es mayor en el grupo de pacientes fumadores que en el de no fumadores, de tal manera que si representamos gráficamente la relación PAS, EDAD en cada uno de los grupos obtenemos unas rectas de regresión como las de la figura



La forma más simple de incorporar la presencia de interacción entre dos variables en una ecuación de regresión consiste en incluir en ésta el producto de ambas:

$$y = b_0 + b_1 \cdot E + b_2 \cdot F + b_3 \cdot E \cdot F$$

donde *E* es la variable edad y *F* fumador (0=no fumador, 1=fumador)

Según esta ecuación el cambio medio de ***y*** cuando la variable ***E*** cambia 1−año es $b_1 + b_3 \cdot F$, es decir que depende también del valor de ***F***, lo que no ocurriría si $b_3=0$.

Selección de variables

Un paso importante en la construcción de un modelo de regresión es el de la **elección de variables** a incluir y cuáles no. Los mecanismos para la selección de variables no son fáciles de especificar ya que dependen en gran medida del tipo de modelo (predictivo o explicativo), del contexto de utilización y de las propias características del proceso analizado. Quizás la única norma clara es que ante dos posibles modelos, similares en otros aspectos, preferiremos el que sea más sencillo y que menos suposiciones necesite para su construcción (es lo que se denomina *principio de parsimonia*).

Para poder decidir entre utilizar un modelo con unas determinadas variables o con otras será preciso disponer de una **medida de comparación entre modelos**.

En la regresión lineal se utiliza para comparar dos modelos la ***F* parcial**, que en el caso de que se contrasten dos modelos que difieren en una sola variable es idéntico a utilizar el valor de la ***t*** para el coeficiente de regresión de la nueva variable.

En la regresión logística, y en general en cualquier modelo de regresión cuyos coeficientes se estimen por el método de máxima verosimilitud, se utiliza el **cociente de verosimilitud**, que es una medida, a partir de los datos de nuestra muestra, de cuánto más probable (verosímil) es un modelo frente al otro. Este parámetro se distribuye según una chi^2 con grados de libertad igual a la diferencia entre el número de variables de los dos modelos. Si no es suficientemente grande decimos que no hay evidencia para pensar que un modelo es mejor

que el otro y por tanto nos quedaremos con el más sencillo.

Existen diferentes **estrategias sistemáticas para la elección de variables** a incluir en los modelos que se van a evaluar. Podemos empezar con un modelo con todas las variables e interacciones −**regresión hacia atrás**−, a partir del cual vamos eliminando variables cuya presencia no mejora la calidad del modelo según el criterio especificado. O por el contrario, podemos empezar con una sola variable independiente e ir añadiendo aquellas variables e interacciones que mejoran significativamente el modelo −**regresión hacia adelante**−. Otra alternativa, no siempre factible si el número de variables es suficientemente grande y no se dispone del software adecuado, es evaluar **todos los modelos de regresión posibles** con todas las combinaciones de variables.

La regresión "**stepwise**", traducida habitualmente como *regresión por pasos*, es una versión modificada del proceso de regresión hacia adelante en la que en cada nuevo paso, cuando se incluye una nueva variable, además se reconsidera el mantener las que ya se había añadido previamente, es decir que no sólo puede entrar una nueva variable en cada paso sino que puede salir alguna de las que ya estaban en la ecuación. El proceso finaliza cuando ninguna variable de las que no están en la ecuación cumple la condición para entrar y de las incorporadas a la ecuación ninguna cumple la condición para salir.

El conjunto de variables que finalmente quede incluido en la ecuación de regresión puede depender del camino seguido a la hora de seleccionarlas, salvo en el caso de que se evaluen todos los modelos de regresión posibles que obviamente sólo tiene una conclusión.

Cualquiera que sea el método que se piense utilizar para la selección de variables éste debe comenzar con un cuidadoso análisis univariante de la posible relación entre la variable dependiente y cada uno de los factores estudiados.

Colinealidad

Algunos autores recomiendan utilizar la estrategia de regresión hacia atrás, comenzando entonces con un modelo en el que se incluyen todas las variables y las posibles interacciones de interés (modelo máximo). Cuando el número de variables es grande con relación al de datos y sobre todo si existe una marcada correlación entre alguna de ellas, puede ocurrir que no sea posible obtener una estimación adecuada de los coeficientes de la ecuación de regresión.

Supongamos, en el caso extremo, que se introduce en la ecuación dos variables que en realidad son la misma, es decir una sola con diferentes nombres. ¿Cómo se reparte entonces el coeficiente de regresión? Si llamamos *X* a esa variable que entra dos veces en la ecuación tendríamos los siguientes términos en la ecuación

$$y = \ldots + b_1 \cdot x + b_2 \cdot x + \ldots$$

o lo que es lo mismo

$$y = \ldots + (b_1 + b_2) \cdot x + \ldots$$

pero ¡hay infinitas formas de repartir una cantidad en dos valores b1 y b2!, por lo que el algoritmo que utiliza el programa de cálculo de los coeficientes de regresión no encuentra una solución.

En el caso de que la relación entre las variables no sea tan perfecta como en el ejemplo planteado, en el que se trata exactamente de la misma, el problema sigue existiendo y aunque quizás el algoritmo de cálculo encuentre una solución para la estimación de los coeficientes puede ocurrir que ésta solución no sea adecuada, debido a un problema de precisión en la estimación, y además siempre será muy dependiente de los datos actuales, de tal manera que una pequeña variación de éstos produce una alteración importante en los valores de los coeficientes de la ecuación. Es lo que en términos matemáticos se conoce como una solución *inestable*.

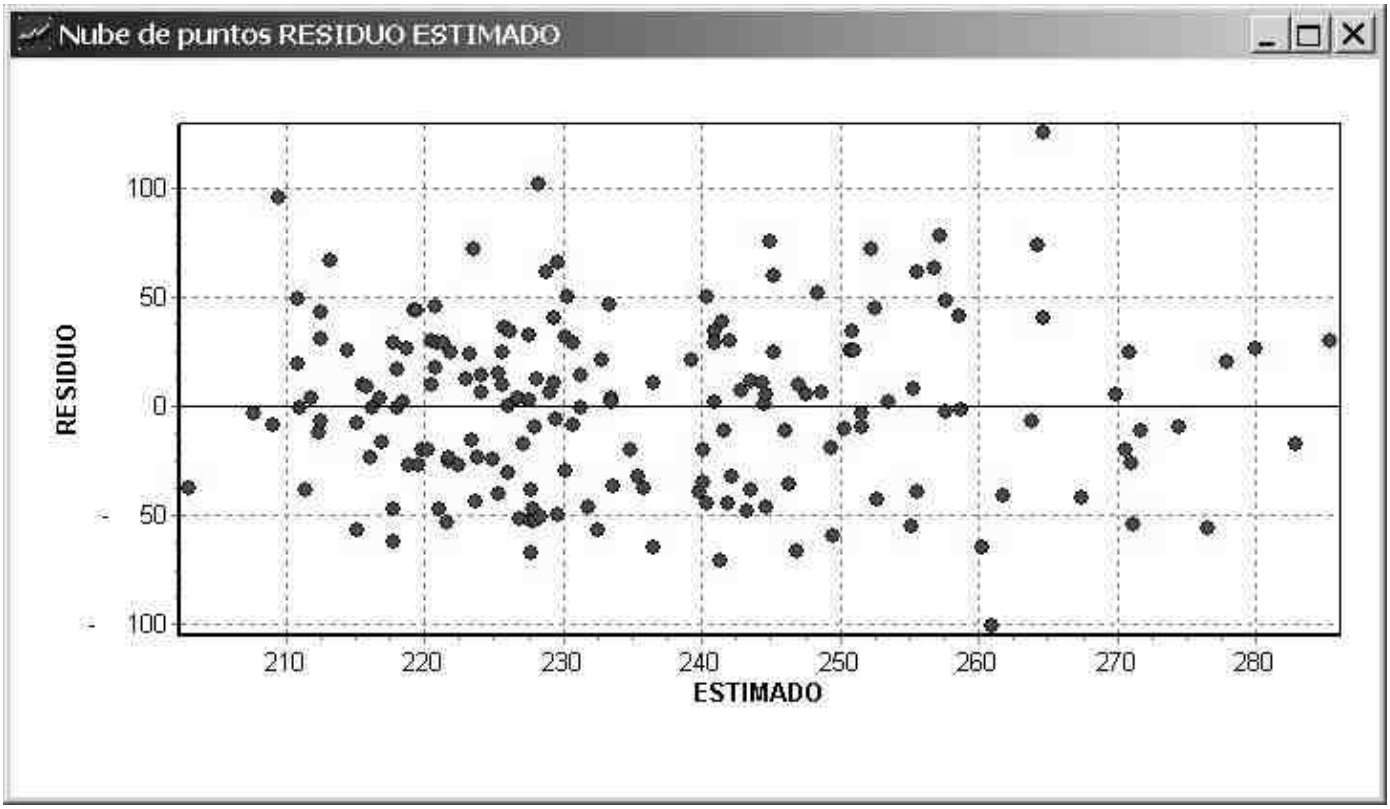
Cuando existe correlación importante entre dos o más variables independientes de una ecuación de regresión se dice en terminología matemática que existe **colinealidad** y es algo que deberíamos comprobar si se produce o no en nuestro modelo de regresión.

Diagnóstico del modelo de regresión

Un aspecto que se olvida frecuentemente es que los modelos de regresión se basan en hacer unas determinadas suposiciones sobre los datos y que éstas no siempre se cumplen, por lo que es preciso comprobar si las hipótesis básicas del modelo se dan en nuestros datos. Es lo que se conoce como **diagnóstico del modelo**.

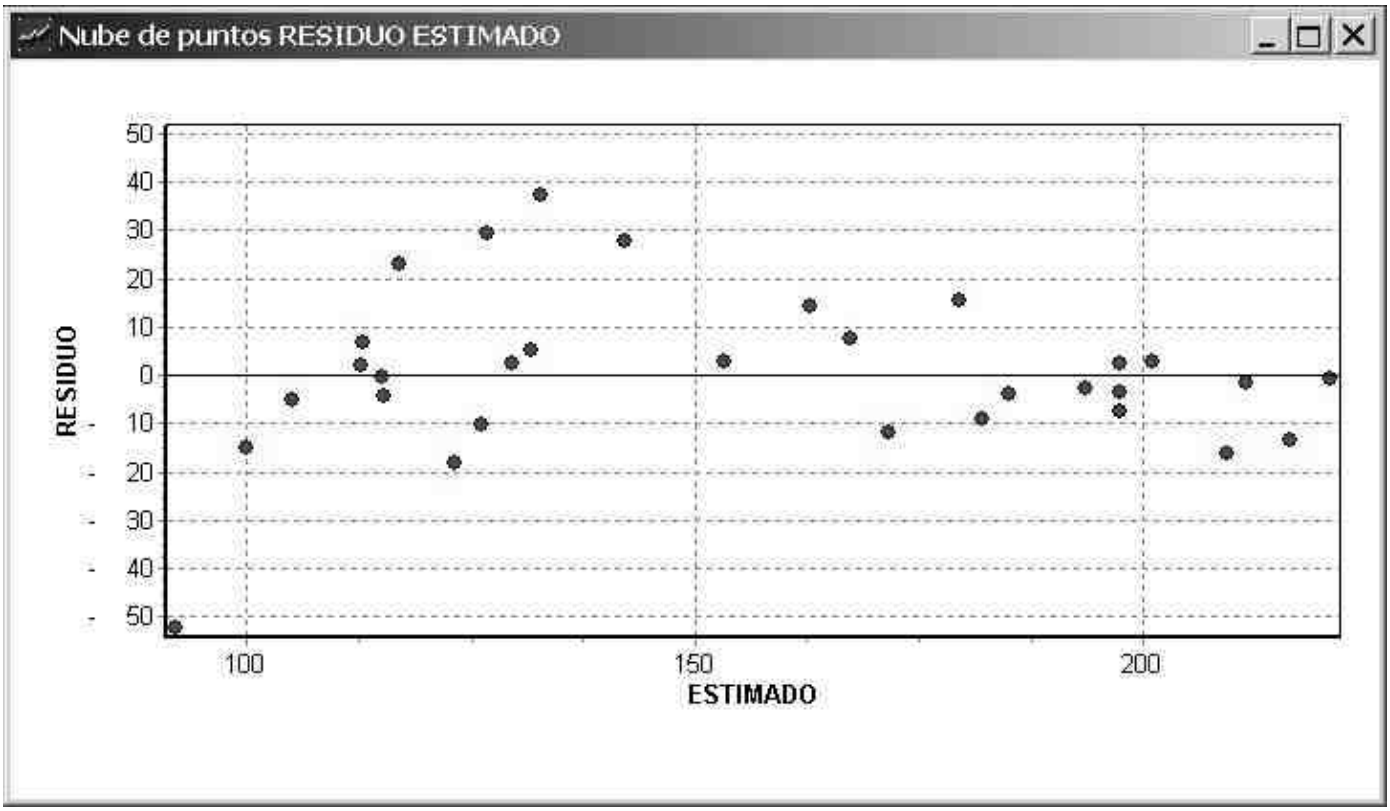
En el caso de los modelos de regresión lineal se utiliza el concepto de **residuo**: diferencia entre el valor observado y el valor estimado por la ecuación de regresión, es decir lo que la ecuación de regresión no explica para cada unidad de observación.

En un modelo de regresión lineal que sea adecuado los residuos deben seguir una distribución normal con media 0 y varianza constante, por lo que un posible diagnóstico puede ser comprobar esa situación. Se puede efectuar de manera formal o mediante una gráfica en la que se representa el valor de los residuos frente al valor estimado, como se ilustra en la siguiente figura



En la gráfica anterior vemos que en este ejemplo efectivamente los residuos se distribuyen de forma simétrica a ambos lados del eje 0 y a lo largo de todo el rango de valores de la estimación y la variabilidad parece constante.

Sin embargo en la siguiente gráfica esto no se cumple lo que en este caso nos está indicando la presencia de un modelo inadecuado



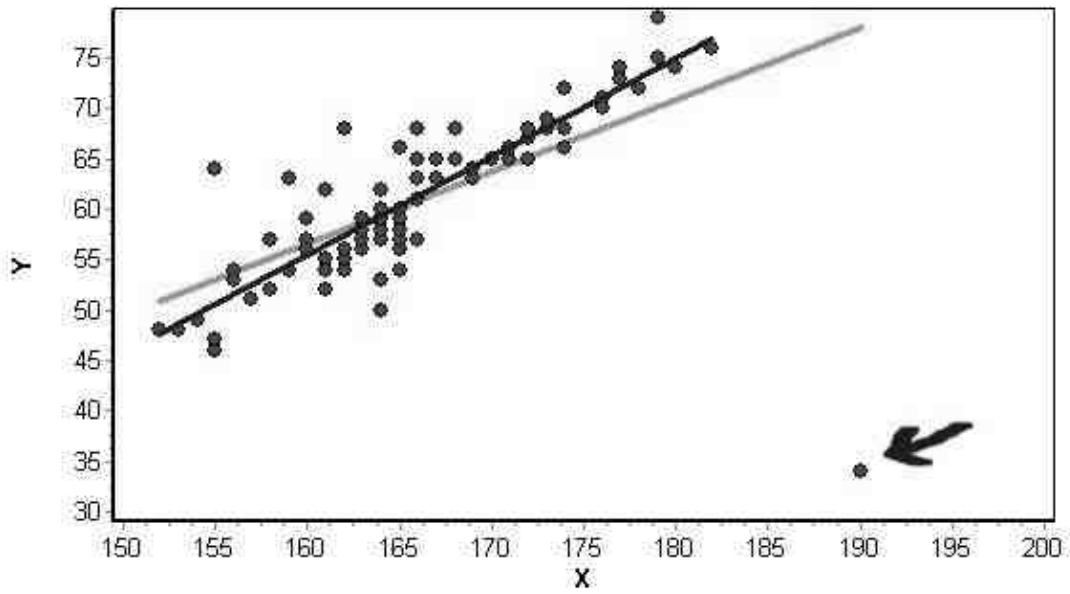
La representación de los residuos frente a cada una de las variables independientes X nos permite detectar la falta de linealidad o la *heterocedasticidad* (se dice que existe heterocedasticidad cuando la dispersión o varianza de la variable no es constante y varía con el valor de ésta). En estos casos puede que sea necesario introducir nuevos términos (como por ejemplo X^2) para considerar esa falta de linealidad, o bien transformaciones matemáticas de las variables.

Para otros tipos de modelos de regresión −regresión logística, modelos de riesgo proporcional de Cox para supervivencia, etc− la metodología es similar pero más compleja.

Valores anómalos

Los **valores extraños (outliers)** son aquellos datos extremos, que parecen anómalos, y que unas veces son debidos a errores de registro al introducir los datos, pero en otras son valores correctos realmente observados. En el caso de la regresión su presencia puede alterar de forma notable los resultados.

En la siguiente figura se representa la recta de regresión (univariante) que se obtiene utilizando todos los datos (color verde) y la que se obtiene cuando se elimina del análisis un sólo dato, el que se señala en la zona inferior derecha. En el primer caso el valor del coeficiente de regresión es 0.98 y en el segundo 0.72. La introducción de ese único dato −en una muestra de 100− produce un cambio en el coeficiente de regresión del 27 %.



Es por tanto muy importante un cuidadoso análisis de los valores extremos e incluso efectuar un análisis de regresión con y sin ellos, para valorar cómo afecta su presencia a los coeficientes de la ecuación de regresión.

Validación del modelo

Los modelos de regresión pueden ser **validados** en otro conjunto de datos de similares características −extraídos de la misma población−, con el fin de evaluar su fiabilidad. Otra posibilidad, cuando se trabaja con muestras grandes, es dividir aleatoriamente la muestra en dos grupos y utilizarlos para obtener dos modelos con el fin compararlos para comprobar si se obtienen similares resultados.

Un índice empleado para validar el modelo se basa en estimar la ecuación de regresión en una de las submuestras y calcular el coeficiente de correlación R_a entre los valores observados y los valores estimados por la ecuación (este coeficiente coincide con el valor del coeficiente de correlación múltiple). Después aplicamos la ecuación de regresión al otro grupo para calcular el valor estimado de *Y* para cada unidad de observación y calculamos el coeficiente de correlación R_b entre ese valor estimado y el valor realmente observado. La diferencia entre el cuadrado de ambos coeficientes $R_a^2 - R_b^2$ se denomina *índice de reducción en la validación cruzada*. Valores de este índice inferiores a 0.1 indican que el modelo es muy fiable mientras que valores superiores a 0.9 corresponden a modelos muy poco fiables.

Presentación de modelos de regresión

De lo anteriormente expuesto parece lógico concluir las siguientes normas de presentación de modelos de regresión

- Indicar en una tabla los coeficientes de la ecuación de regresión, con su error estándar, estadístico de contraste para el coeficiente (t, chi^2, F, test de Wald) y valor de probabilidad asociado.
- Especificar qué variables fueron candidatas a ser consideradas en la ecuación de regresión y qué camino se siguió para seleccionar las definitivamente incluidas
- Especificar si se evaluó la posible presencia de interacción entre las variables
- Especificar si se comprobó la posible existencia de colinealidad entre variables.
- Especificar si se revisaron los valores extremos y si éstos se incluyeron en el modelo o no, y cómo afectan a los resultados.
- Especificar qué diagnósticos se han realizado sobre el modelo.

- Especificar si se efectuó algún tipo de validación del modelo

---

Para finalizar no nos resistimos a citar un texto de KJ Rothman, que aunque un poco largo no puede ser más claro y cuya opinión, aunque un poco extrema, nos parece digna de tenerse en cuenta:

"La primera experiencia que se tiene con el análisis multivariado le deja a uno con la impresión de que acaba de serle revelado un milagro de la tecnología del análisis de datos; el método permite controlar la confusión y evaluar las interacciones de multitud de variables con gran eficiencia estadística. Mejor aún, un ordenador efectua todos los cálculos y te imprime con limpieza los resultados. La temeraria experiencia de solicitarle que consiga todas estas metas analíticas, para luego simplemente poner en orden y publicar la sofisticada salida con apenas una pausa para volver a teclear, es indiscutiblemente tentadora. Ciertamente puede incluso llegar a ser decepcionante ver cómo el análisis que culmina el trabajo de semanas, meses o años de recogida de datos se acaba en un tiempo tan corto y los resultados se comprimen de manera tan compacta.

Por útil que pueda ser sin embargo, el análisis multivariado no es una panacea estadística. Su mayor inconveniente radica en la barrera que inserta entre el investigador y los datos. Otros métodos analíticos facilitan una comprensión íntima de éstos, haciendo consciente al investigador de la existencia de irregularidades o deficiencias −unas pocas entrada de celda críticas con frecuencias pequeñas, por ejemplo−. Los métodos multivariados dificultan esta intimidad con los datos. Otro inconveniente, relacionado con éste, se halla en la falta de capacidad para comunicar a otros el mensaje de los resultados. Algunos lectores se sienten poco familiarizados e incómodos con los modelos matemáticos que se maneja y todos ellos, lo mismo que los investigadores, obtienen una comprensión más clara de los datos si lo que se presentan son frecuencias tabuladas.

La creciente disponibilidad de hardware y software ha significado el pistoletazo de salida para una avalancha de datos mal digeridos, gran parte de ella caracterizada por su análisis desestructurados y pobremente interpretados, que conducen al investigador hacia la meta de la investigación por accidente, si es que le conducen a alguna parte. En la literatura científica, uno se encuentra frecuentemente con situaciones en las que los métodos analíticos sencillos podrían haber sido aplicados, pero se los dejó de lado en favor del análisis multivariado sin una razón clara. Al epidemiólogo, de entrada, le merece más la pena apoyarse, siempre que sea posible, en los procedimientos del análisis estratificado, que son más directos y que engendran una mayor familiaridad con los datos tanto para el investigador como para el lector."

Y como complemento a este texto recomendamos la lectura del artículo Commentary: Prognostic models: clinically useful or quickly forgotten? donde muy lúcidamente se exponen las posibles razones de por qué la gran mayoría de los modelos predictivos que se publican en medicina tienen una vida tan efímera.

Enlaces de interés

- ***Commentary: Prognostic models: clinically useful or quickly forgotten?***
  Jeremy C Wyatt and Douglas G Altman
  BMJ 1995; 311: 1539−1541. [Texto completo]
- Rice Virtual Lab in Statistics
  Case Studies
  Examples of real data with analyses and interpretation
- Multiple regression: basic concepts and procedures
  Department of psychology.
  University of Exeter
- Multiple regression. Selecting the best equation (PDF)

Department of Mathematics and Statistics.
University of Saskatchewan, Saskatoon, Saskatchewan Canada

- Selection of the best regression model (PDF)
  Biostatistics.
  Harvard School of Public Health

Bibliografía seleccionada

- **Applied regression analysis and other multivariate methods**
  David G. Kleinbaum, Lawrence L. Kupper, Keith E. Muller, Azhar Nizam
  Kupper, Muller, Nizam
  Ed. Duxbury Press 1998
- **Applied Logistic Regression**
  David W. Hosmer, Stanley Lemeshow
  Ed.John Wiley
  New York 1989
- **Epidemiological research**
  David G. Kleinbaum, Lawrence L. Kupper, Hal Morgentern
  Ed.John Wiley
  New York 1982
- **Epidemiología moderna**
  Kenneth J. Rothman
  Ed.Díaz de Santos
  Madrid 1986

Algunos ejemplos de artículos sobre hipertensión en los que se utiliza modelos de regresión

- ***Survival in treated hypertension: follow up study after two decades***
  Ove K Andersson, Torbjörn Almgren, Bengt Persson, Ola Samuelsson, Thomas Hedner, and Lars Wilhelmsen
  BMJ 1998; 317: 167−171. [Abstract] [Texto completo]
- ***Case−control study of stroke and the quality of hypertension control in north west England***
  Xianglin Du, Kennedy Cruickshank, Roseanne McNamee, Mohamad Saraee, Joan Sourbutts, Alison Summers, Nick Roberts, Elizabeth Walton, and Stephen Holmes
  BMJ 1997; 314: 272. [Abstract] [Texto completo]
- ***Obstructive sleep apnoea syndrome as a risk factor for hypertension: population study***
  Peretz Lavie, Paula Herer, and Victor Hoffstein
  BMJ 2000; 320: 479−482. [Abstract] [Texto completo]
- ***Diabetes mellitus and raised serum triglyceride concentration in treated hypertension−−are they of prognostic importance? Observational study***
  Ola Samuelsson, Kjell Pennert, Ove Andersson, Goran Berglund, Thomas Hedner, Bengt Persson, Hans Wedel, and Lars Wilhelmsen
  BMJ 1996; 313: 660−663. [Abstract] [Texto completo]
- ***Pressor reactions to psychological stress and prediction of future blood pressure: data from the Whitehall II study***
  Douglas Carroll, George Davey Smith, David Sheffield, Martin J Shipley, and Michael G Marmot
  BMJ 1995; 310: 771−775. [Abstract] [Texto completo]

- ***Obesity, Hypertension, and the Risk of Kidney Cancer in Men***
  Chow W.−H., Gridley G., Fraumeni J. F., Järvholm B.
  [Abstract] [Texto completo]
  N Engl J Med 2000; 343:1305−1311, Nov 2, 2000. **Original Articles**
- ***The Effect of Nisoldipine as Compared with Enalapril on Cardiovascular Outcomes in Patients with Non−Insulin−Dependent Diabetes and Hypertension***
  Estacio R. O., Jeffers B. W., Hiatt W. R., Biggerstaff S. L., Gifford N., Schrier R. W.
  [Abstract] [Texto completo]
  N Engl J Med 1998; 338:645−652, Mar 5, 1998. **Original Articles**

Indice de artículos

Principio de la página