COLABORACIÓN ESPECIAL

DATOS ANÓMALOS Y REGRESIÓN LOGÍSTICA ROBUSTA EN CIENCIAS DE LA SALUD

Francisco Cutanda Henríquez

Hospital General Universitario Gregorio Marañón.

RESUMEN

La regresión logística tiene numerosas aplicaciones en Ciencias de la Salud. Existe una amplia literatura respecto a los métodos a seguir y al modo de hallar los estimadores de los parámetros a partir de las observaciones. Estos métodos están incorporados en todos los paquetes estadísticos usuales. Los estimadores son los llamados de "máxima verosimilitud", es decir, son aquéllos que hacen que las observaciones obtenidas sean las más probables entre todos los posibles modelos que pudiéramos utilizar. Las buenas propiedades de los estimadores de máxima verosimilitud están ampliamente demostradas.

Sin embargo, en la práctica existe una serie de circunstancias que pueden ocasionar la aparición de "datos anómalos", es decir, observaciones que no corresponden al modelo logístico que utilizamos como hipótesis. En ocasiones, estas observaciones anómalas pueden tener un fuerte efecto sobre el ajuste y, por tanto, llevarnos a una conclusión equivocada. Las causas de estos datos anómalos dependen mucho del estudio en cuestión, pero pueden señalarse errores de clasificación, observaciones (sujetos) con características especiales que se han pasado por alto, incertidumbres en la medida de algunos parámetros, etc.

El problema de los estimadores de máxima verosimilitud es que no son "robustos", es decir, su sensibilidad a datos anómalos puede ser arbitrariamente grande, y una minoría de datos anómalos puede dar lugar a un modelo logístico erróneo. En este trabajo expondremos dos casos que ilustran las posibles consecuencias, y discutiremos la aplicación de métodos robustos.

Palabras clave: Bioestadística. Regresión logística. Probabilidad.

ABSTRACT

Outliers and Robust Logistic Regression in Health Sciences

Logistic regression methods have many applications in Health Sciences. There is a vast literature about procedures to be followed and the way to find the estimators for the parameters from the observed values, and these methods are implemented to all the usual statistical packages. These estimators are of the "maximum likelihood" kind, i.e., they are the ones that make the observed values the most probable among all the models that could have been used. The good properties of the maximum likelihood estimators are widely demonstrated.

However, there are some practical circumstances that may cause the presence of "outliers", i.e., observed values not corresponding to the logistic model we are assuming as a hypothesis. Occasionally, these anomalous observations can have a strong effect on the fit, and lead the study to the wrong conclusion. The causes of these outliers depend on the particular study, but it is possible to point out classification errors, observations (subjects) with special features which have not been taken into account, uncertainty in the measurement of some parameters, etc.

The problem with maximum likelihood estimators is that they are not "robust", i.e., their sensitivity to outliers could be arbitrarily large, and a minority of outliers could lead to a wrong logistic model. In this work, we will show two cases illustrating possible consequences, and we will discuss the application of robust methods.

Keywords: Biostatitics. Logistic models. Regression analysis. Probability.

Correspondencia:
Hospital General Universitario Gregorio Marañón
Calle Dr. Esquerdo, 46
28007 MADRID
Tel: 914265129
Correo electrónico: francisco.cutanda@salud.madrid.org

INTRODUCCIÓN

El uso de procedimientos de regresión está extendido en Ciencias Económicas, Sociales, Experimentales y de la Salud. En el método de regresión logística se pretende llegar a un modelo que explique una característica binaria (un "sí" o "no", "éxito" frente a "fracaso", "enfermo" frente a "sano", etc.) y su probabilidad a partir de cierto número de variables observadas. Por ejemplo, un estudio epidemiológico puede recabar información sobre un número grande de sujetos, su sexo, edad, si fuma o no, cuánto ejercicio hace, ocupación, lugar de residencia, estado civil, etc. además de si han padecido o no infarto. La regresión logística ayudaría a descartar cuáles de estas variables son realmente "explicativas" de la probabilidad de infarto y, para las que son explicativas, produciría un modelo matemático ajustado a nuestras observaciones que podría hacer predicciones. A veces el primer aspecto, poder descartar el efecto de una variable, es más importante incluso que el segundo.

En Ciencias de la Salud, particularmente en Epidemiología, se plantean usualmente problemas de alta complejidad: de cada individuo son recogidas múltiples variables, las muestras son muy grandes. Para realizar el estudio estadístico es necesario el uso de bases de datos y paquetes estadísticos. La teoría, sin embargo, es relativamente sencilla y descansa sobre un principio universal: "el modelo que mejor ajusta las observaciones es aquél que hace más probable la muestra obtenida de entre todos los modelos posibles". Este es el principio de "máxima verosimilitud", y los métodos de ajuste a modelos logísticos, lineales, de Poisson u otros se basan en este principio[1,2,3].

Un dato anómalo, a veces conocido por el nombre inglés "outlier", es una observación de la muestra que en realidad no corresponde al modelo buscado[4,5]. Supongamos que ha habido un error de trascripción, y un paciente que sí padeció un infarto fue clasificado como que no lo padeció, y que en otro paciente ocurrió el error inverso. Si el haber padecido infarto es un factor importante, está claro que estos dos resultados, tal como han llegado a nuestras manos, no pueden ser explicados por el mismo modelo que los demás. Un error de transcripción en una variable con poca significación puede no tener consecuencias, pero dependiendo de la naturaleza del experimento el caso podría ser otro.

Otra situación que introduciría datos anómalos en el experimento sería, por ejemplo, que no se hubieran incluido datos referentes a la dieta, y que existiera algún sujeto vegetariano, para el que la incidencia de infarto sería totalmente distinta.

Otras causas que harían anómalo un dato pueden ser más sutiles. Quizá en la muestra de sujetos en estudio cuyos datos estamos analizando hay ciertas correlaciones entre el sexo, ser fumador y la edad, de modo que una anciana fumadora sea un caso aislado. Esto es lo que llamaríamos un dato "extremo" puesto que, si pudiéramos hacer una gráfica de las variables que estamos considerando, quedaría apartado de los demás.

La inclusión de un dato anómalo no tiene por qué cambiar el resultado apreciablemente respecto al caso en que este dato no se dio, pero en ocasiones sí ocurre, y el problema que se nos plantea es que en un problema grande, con muchas variables y muchos sujetos, los datos anómalos pasan desapercibidos.

Se dice que un estimador es "robusto" cuando es poco sensible a la presencia de datos anómalos en la muestra. Es fácil demostrar que los estimadores de máxima verosimilitud no son robustos, a pesar de poseer otras propiedades deseables.

Para ilustrar cuál puede ser la diferencia entre un estimador robusto y uno que no lo es pensemos en el siguiente experimento sencillo. Imaginemos que tenemos una muestra de 10 niños de cinco años de edad y queremos hallar la media y la mediana de su estatura. Supongamos que uno de los sujetos ha sido introducido erróneamente como de cinco años pero no sabemos su edad. Si las estaturas de los niños (en m.) son 1,10, 1,05, 1,15, 1,20, 1,16, 1,11, 1,06, 1,13, 1,14, 1,15, la media será 1,125 y la mediana será 1,135. Si el último niño de la lista, cuya edad no conocemos hubiera resultado medir 1,65, la media hubiera sido 1,170 y la mediana 1,135. Si por el contrario hubiera sido un bebé de 0,65 m la media sería 1,075 y la mediana sería 1,12. La media es un estadístico no robusto, y eso se ve en la variación que un único dato anómalo puede producir. Esta variación podría ser tan grande como se quisiera. La mediana es un estadístico robusto, para el que la anomalía produce un efecto limitado.

Este es un ejemplo muy simple, porque el dato anómalo puede reconocerse a simple vista y eliminarse. En problemas complejos esto no es posible, como en la regresión logística, a menos que recurramos a técnicas de diagnóstico especiales para identificar los datos anómalos. Veremos en este trabajo cómo estas técnicas tienen una utilidad limitada, y puede resultar recomendable utilizar un método robusto.

MATERIAL Y MÉTODOS

En este trabajo se han estudiado dos conjuntos de datos proporcionados por el Hospital General Gregorio Marañón.

1. GM1. Mortalidad en UCI pediátrica. La variable dependiente en este estudio es la probabilidad de muerte en pacientes pediátricos tras una parada cardiorrespiratoria. Consta de 147 casos y se estudian siete variables de tipo categórico, con 7, 7, 2, 2, 2, 2 y 6 niveles respectivamente.

2. GM2. Fibrosis en pacientes coinfectados con VHC y VIH. La variable en estudio es la probabilidad de aparición de estadíos avanzados de fibrosis (F3-F4). El estudio incluye a 220 pacientes y se han recogidos datos muy diversos como covariables, sumando cincuenta y una variables dicotómicas y continuas.

Se ha utilizado el paquete estadístico R para realizar los estudios[6,7]. En ambos casos, se ha realizado una primera regresión logística con todas las variables y hallado los valores de significación de cada una de las variables. Se han eliminado todas las variables no significativas al 5%, lo cual ha dejado cuatro variables en ambos casos. A partir de ese punto hemos realizado primero un estudio por métodos clásicos, aplicando un método diagnóstico, y el mismo estudio por métodos robustos.

Como método diagnóstico presentaremos tanto el estadístico de Cook[8,9] como el de Lee[10]. El primero da, para cada observación o sujeto, la magnitud del cambio del ajuste cuando se prescinde de esa observación o sujeto. El estadístico de Lee da el cambio que la eliminación de ese sujeto produciría en la significación de una variable para cada variable y cada sujeto. Estos dos métodos diagnósticos deberían dar valores uniformes para todos los sujetos. Si existe un sujeto que da valores muy diferentes a los de los demás, es probable que nos encontremos con un dato anómalo influyente.

El estadístico de Cook mediría su influencia global sobre el ajuste, y el de Lee su influencia sobre la significación de cada variable. Si el estadístico de Lee para una de las variables es uniforme para cada sujeto, podemos estar seguros de que ningún

dato anómalo va a favorecer o desfavorecer la significación de esta variable, llevándonos a descartarla o aceptarla erróneamente. Se puede dar el caso de existir dos "outliers", ambos detectables por el estadístico de Cook que según el estadístico de Lee tendrían efecto sobre la significación de una variable cada uno, siendo intrascendentes para las demás.

El método robusto que vamos a aplicar es el propuesto por Cantoni y Ronchetti[11]. Este método no es específico de regresión logística, sino de modelos lineales generalizados, a los que pertenece la regresión logística[12]. La ventaja que ofrece es que proporciona un método para la selección de variables, aparte de obtener el ajuste. Como en la mayoría de métodos robustos se trata de rebajar la verosimilitud asociada a sujetos muy influyentes en el ajuste o a sujetos extremos. Los estimadores que proponen son de tipo Mallows, y para su cálculo proporcionan funciones para S-plus, adaptables a R[13].

Es importante recordar que un estadístico robusto obtenido por este método puede carecer de algunas de las propiedades de los estadísticos de máxima verosimilitud. En todo caso, usualmente se obtienen estimadores con mayor desviación estándar con el método robusto. Es el precio a pagar a cambio del comportamiento robusto.

RESULTADOS

El modelo de regresión logística que queremos ajustar se escribe como:

$$GM1 \quad log\left(\frac{p_1}{1 - p_1}\right) = a_0 + a_1x_1 + a_2x_2 + a_3x_3\ K...$$

$$GM2 \quad log\left(\frac{p_2}{1 - p_2}\right) = b_0 + b_1y_1 + b_2y_2 + b_3y_3\ K...$$

donde p_1, p_2 es la probabilidad del evento en estudio, a_0, a_1,... b_0, b_1,... son coeficientes a determinar mediante el ajuste y x_0, x_1,.. y_0, y_1,.... son las covariables, cuyos valores son 0 y 1 para las de tipo dicotómico.

1. GM1. Mortalidad en UCI pediátrica

a) Estudio clásico. Los estimadores de máxima verosimilitud para el modelo de cuatro variables que resulta de eliminar las menos significativas aparecen en la tabla 1 (modelo GM1CL1).

La significación es la probabilidad de que el coeficiente sea cero, es decir, que la covariable correspondiente pueda ignorarse. Vemos que los valores sugieren que los coeficientes a_1 y a_2 podrían ser prescindibles. Esto nos conduce al modelo GM1CL2 de la tabla 2.

Estos dos modelos son plenamente aceptables. Los estadísticos de bondad del ajuste son buenos y nada hace pensar que la conclusión (la significación de las variables a_3 y a_4) sea discutible. Sin embargo, al observar el estadístico de Cook para cada sujeto, encontramos que los sujetos 58 y 78

Tabla 1

Estimadores, desviaciones típicas y significaciones para el modelo clásico de cuatro variables del problema GM1

GM1CL1	Estimador	Desv. Est.	Significación
a_0	1,7989	0,3459	0,0000002
a_1	-3,0895	1,4593	0,03426
a_2	2,4934	1,0957	0,02286
a_3	-2,3189	0,4639	0,0000006
a_4	-1,3626	0,4477	0,00234

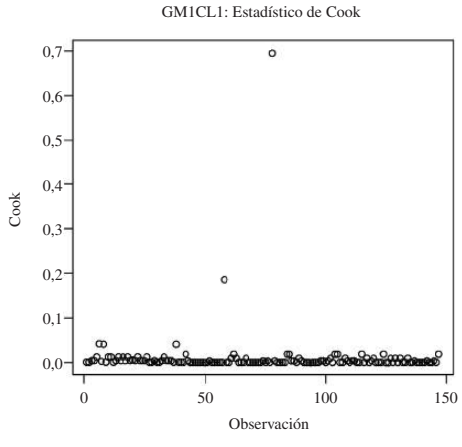
Tabla 2

Estimadores, desviaciones típicas y significaciones para el modelo clásico de dos variables del problema GM1

GM1CL2	Estimador	Desv. Est.	Significación
a_0	1,8552	0,3391	0,00000004
a_3	-2,4037	0,4466	0,00000007
a_4	-1,0712	0,4208	0,0109

Figura 1

Estadístico de Cook para el modelo GM1CL1



son claramente discrepantes de los demás, es decir, están ejerciendo una influencia que hace variar el ajuste del que se obtendría de las demás variables (figura 1). Si eliminamos estas dos variables y ajustamos de nuevo, obtenemos el modelo GM1modCL1 que se muestra en la tabla 3.

Las variables a_1 y a_2 tienen ahora una significación despreciable. Al eliminarlas obtenemos el modelo GM1modCL2 (tabla 4).

Vemos que la gráfica de Cook para este modelo presenta pocas diferencias. Para verificar que estos dos sujetos son datos anormalmente influyentes en estas variables, obtenemos el estadístico de Lee para cada una de las cuatro variables (Figura 2), donde podemos comprobar que estos dos sujetos (58 y 78) tienen realmente una influencia anómala apareciendo con valores destacados en las gráficas de estas dos variables.

b) Estudio robusto. El primer ajuste robusto (GM1CR1) se muestra en la tabla 5 que de nuevo sugiere la eliminación de las dos primeras variables resultando el modelo GM1CR2 de la tabla 6.

2. Fibrosis en pacientes coinfectados por VIH y VHC

a) Estudio clásico. Los estimadores de máxima verosimilitud y la significación de las variables que se obtienen, una vez eliminadas aquéllas no significativas al 0.05, es el que se muestra en la tabla 7 (GM2CL1):

Tabla 3

Estimadores, desviaciones típicas y significaciones para el modelo clásico de cuatro variables del problema GM1, con los datos 58 y 78 eliminados

GM1modCL1	Estimador	Desv. Est.	Significación
a_0	1,6936	0,3387	0,0000006
a_1	-17,4960	51,6901	0,7350
a_2	8,9830	16,7989	0,5928
a_3	-2,2222	0,4182	0,000002
a_4	-1,2870	0,4470	0,00397

Tabla 4

Estimadores, desviaciones típicas y significaciones para el modelo clásico de dos variables del problema GM1, con los datos 58 y 78 eliminados

GM1modCL2	Estimador	Desv. Est.	Significación
a_0	1,8514	0,3406	0,00000005
a_3	-2,4177	0,4465	0,00000006
a_4	-1,0025	0,4468	0,0180

Tabla 5

Estimadores, desviaciones típicas y significaciones para el modelo robusto de cuatro variables del problema GM1

GM1CR1	Estimador	Desv. Est.	Significación
a_0	1,8276	0,3581	
a_1	-2,6953	1,4085	0,05568
a_2	1,9266	0,9859	0,05069
a_3	-2,3290	0,4738	0,0000009
a_4	-1,4065	0,4593	0,00220

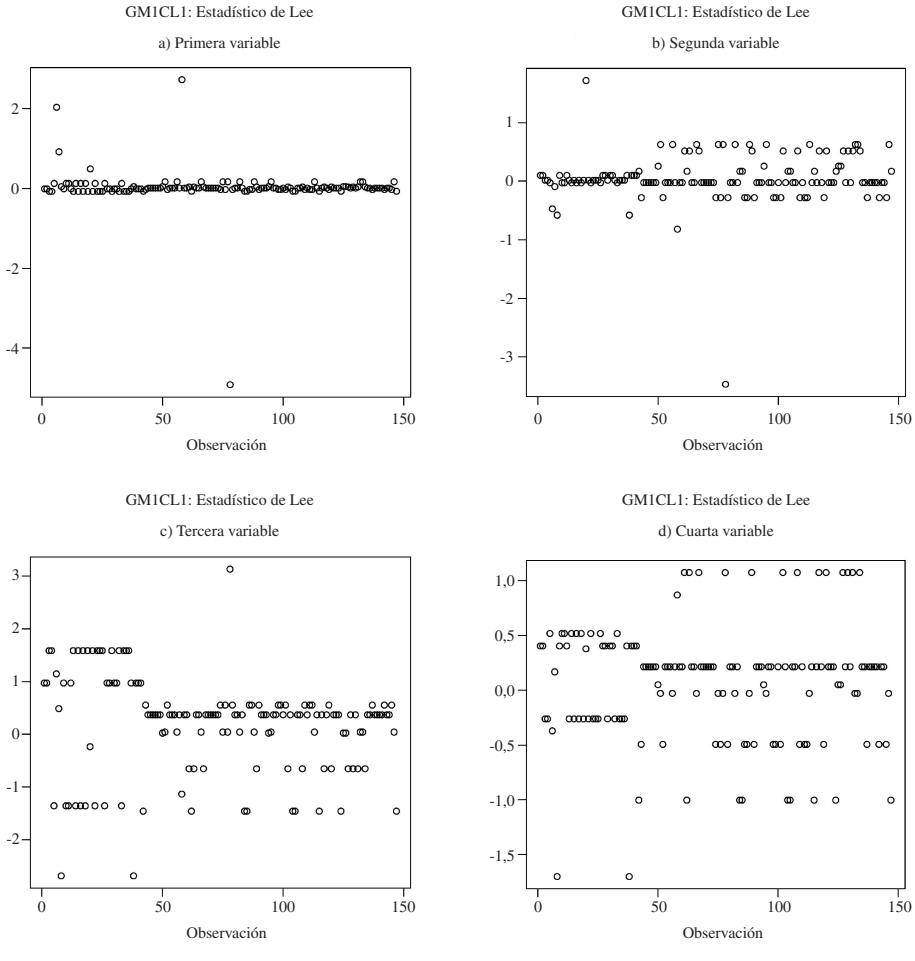
Tabla 6

Estimadores, desviaciones típicas y significaciones para el modelo robusto de dos variables del problema GM1

GM1CR2	Estimador	Desv. Est.	Significación
a_0	1,8707	0,3497	
a_3	-2,3958	0,4510	0,0000001
a_4	-1,0749	0,4308	0,0126

Figura 2

Estadístico de Lee para el modelo GM1CL1



Eliminando la primera variable, de escasa significación, resulta el modelo GM2CL2 de la tabla 8.

Tabla 7

Estimadores, desviaciones típicas y significaciones para el modelo clásico de tres variables del problema GM2

GM2CL1	Estimador	Desv. Est.	Significación
b_0	-3,9475	1,7569	0,0246
b_1	0,001632	0,001568	0,2981
b_2	4,6095	1,4649	0,001652
b_3	-0,01254	0,003536	0,00039

En este caso, la gráfica del estadístico de Cook y las de Lee (Figuras 3 y 4) muestran que el sujeto 50 es altamente anómalo. Si repetimos los ajustes eliminando esta obser-

Tabla 8

Estimadores, desviaciones típicas y significaciones para el modelo clásico de dos variables del problema GM2

GM2CL2	Estimador	Desv. Est.	Significación
b_0	1,8707	0,3497	
b_2	-2,3958	0,4510	0,0000001
b_3	-1,0749	0,4308	0,0126

Figura 3

Estadístico de Cook para el modelo GM2CL1

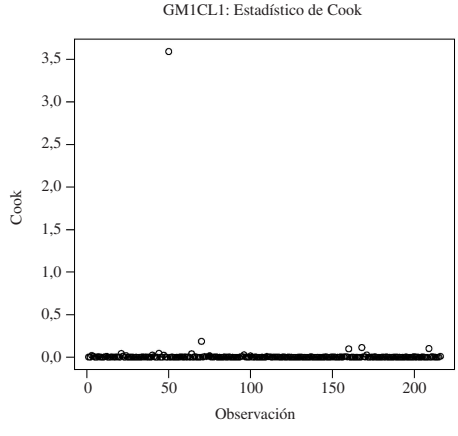


Tabla 9

Estimadores, desviaciones típicas y significaciones para el modelo clásico de tres variables del problema GM2, con el dato 50 eliminado

GM2modCL1	Estimador	Desv. Est.	Significación
b_0	-6,4972	1,9587	0,000909
b_1	0,009010	0,002233	0,00005
b_2	5,8687	1,6138	0,000276
b_3	-0,01192	0,003711	0,001319

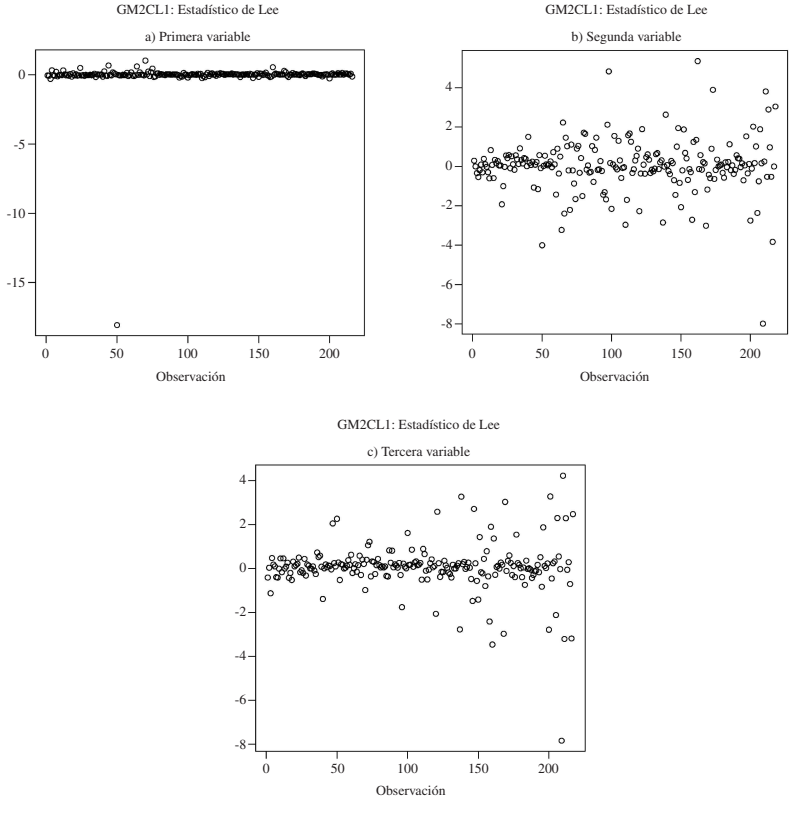
Tabla 10

Estimadores, desviaciones típicas y significaciones para el modelo robusto de tres variables del problema GM2

GM2CR1	Estimador	Desv. Est.	Significación
b_0	-6,6293	2,3622	
b_1	0,008635	0,002764	0,001787
b_2	6,7271	1,9888	0,007184
b_3	-0,01505	0,004880	0,00204

Figura 4

Estadístico de Lee para el modelo GM2CL1



vación obtenemos el modelo GM2modCL1 de la tabla 9.

Y ahora, curiosamente, las tres variables tienen buena significación. La presencia del dato anómalo nos forzó a ignorar una variable importante.

b) Estudio robusto. El modelo robusto de tres variables obtenido mediante el método de Cantoni y Ronchetti es (GM2CR1) el de la tabla 10.

Con este método hemos alcanzado la conclusión correcta, (la dada por el modelo GM2modCL1) sin utilizar diagnósticos ni eliminar observaciones. Es decir, el estimador de Cantoni y Ronchetti censura con éxito un dato claramente anómalo.

COMENTARIOS

En el primero de los casos expuestos, dos datos se han manifestado como anómalos a través de los métodos diagnósticos utilizados, y el estadístico de Lee ha señalado que las variables primera y segunda serían las afectadas por estas anomalías. En efecto, vemos que ha cambiado drásticamente la significación de estas dos variables cuando han sido ignoradas, y que si hubiéramos utilizado desde el primer momento el método robusto éstas habrían sido señaladas como poco significativas al 5% de significación.

En el segundo caso, una situación un tanto inversa se ha podido ver: aquélla en que la presencia del dato anómalo ha quitado significación a una de las variables, invitando al experimentador a eliminarla del modelo. El ajuste con el dato eliminado y el ajuste robusto apuntan ambos a un modelo con tres variables.

Hay que destacar que las variables dependientes en estudio, la mortalidad y el estadio de fibrosis, tienen distinta naturaleza. No hay incertidumbre en la determinación de la mortalidad, mientras que la evaluación del estadío de la fibrosis está sujeta a una cierta subjetividad, al tratarse de un fenómeno en que una variedad amplia en la gravedad ha sido forzada a encajar en cinco categorías. Es de suponer que hay más pacientes dudosos en el estudio, cuya fibrosis ha sido clasificada como más o menos avanzada de lo que realmente es, aunque, al ser casos frontera, no aparezcan como "outliers" muy influyentes.

Hemos visto cómo la aplicación de un procedimiento "clásico", de máxima verosimilitud, que será el que nos proporcione la mayoría de programas informáticos estadísticos por defecto, puede dar un resultado impecable, con variables altamente significativas, un buen valor para la bondad del ajuste, y, sin embargo, llevarnos a una conclusión errónea, no sólo en la estimación de los parámetros, sino también en la selección de las variables significativas.

AGRADECIMIENTOS

El autor desea expresar sus agradecimientos a Silvia Vargas Castrillón, José María Bellón y Alfonso García Pérez.

BIBLIOGRAFÍA

1. Vélez Ibarrola, R, García Pérez, A. Principios de inferencia estadística. Madrid:UNED; 1993.

2. Mould RF. Introductory Medical Statistics. Bristol:Adam Hilger; 1989.

3. García Pérez A. Métodos Avanzados de Estadística Aplicada. Técnicas Avanzadas. Editorial UNED, 2005.

4. García Pérez A. Técnicas actuales de Estadística aplicada. Madrid:UNED; 2006.

5. García Pérez A. Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo. Madrid:UNED; 2005.

6. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2007.

7. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.2-30. 2007.

8. Cook RD. Detection of influential observation in linear regression. Technometrics. 19(1):15. Feb 1977.

9. Cook RD. Influential observations in linear regression. J Am Stat Assoc. 1979; 74(365):169.

10. Lee AH. Assessing partial influence in generalized linear models. Biometrics, 44(1):71. 1988.

11. Cantoni E, Ronchetti, E. Robust inference for generalized linear models. J Am Stat Assoc. 2001; 96(455):1022.

12. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc [Ser A]. 1972; 135(3):370-384.

13. Cantoni E. Analysis of robust quasi-deviances for generalized linear modelsJ Stat Softw. 2004; 10(4), 2004.