EVALUAR

Laboratorio de Evaluación Psicológica y Educativa

Evaluar, 6 (2006), 52 – 67
ISSN 1667-4545

ARTICULO METODOLÓGICO

Fundamentos del Análisis de Regresión Logística en la Investigación Psicológica

Ana María Alderete[1]*

* Universidad Nacional de Córdoba

Resumen. En este artículo se presenta el modelo de regresión logística, apropiado para predecir una variable categórica dicotómica. Esta situación es muy común en la investigación psicológica y en otras disciplinas de las ciencias sociales. Este modelo tiene ventajas sobre el análisis discriminante al no requerir supuestos como el de normalidad o de homocedasticidad, que en muchos casos son difíciles de cumplir. Por otro lado, la regresión logística tiene semejanzas con la regresión múltiple: cuenta con contrastes estadísticos, puede incorporar efectos no lineales y permite realizar diversos diagnósticos. En la actualidad este método es ampliamente utilizado tanto en estudios observacionales como de encuesta y experimentales. Se presentan de manera sencilla los fundamentos lógicos y estadísticos del método, ilustrándose con un ejemplo los procedimientos de cálculo e interpretación de los resultados.

Palabras clave: regresión logística - análisis multivariado - datos categóricos

1. Introducción

Cuando una pregunta de investigación se orienta a conocer cómo inciden varias variables independientes sobre una variable dependiente, el modelo de regresión lineal clásico permite encontrar una respuesta. En ese caso no se presentan restricciones respecto de las variables independientes o predictoras, pero sí respecto de la variable dependiente que se supone debe ser continua y medida al menos en un nivel intervalar. Sin embargo, en muchas circunstancias de investigación en psicología u otras disciplinas de las ciencias sociales la

[1] Por favor dirigir la correspondencia relacionada con este artículo a:
Ana María Alderete
Lic. en Psicología – Profesora Titular de la Cátedra de Metodología de Investigación,
Facultad de Psicología, UNC
Dirección : Manuel Carlés 3688 CP 5008, Córdoba, Argentina
Teléfono: (351) 4767924
E-mail: aldereteanam@yahoo.com.ar

variable dependiente es categórica (dicotómica o politómica). Por ejemplo, cuando evaluamos estado nutricional, *eutrófico* o *desnutrido*; o niños con problemas de aprendizaje, *con problemas* de aprendizaje y *sin problemas*. En estos casos es apropiado el método de regresión logística, que se caracteriza por ser un modelo de probabilidad lineal.

Este modelo es una generalización del modelo de regresión lineal clásico para variables dependientes categóricas dicotómicas (Ato y García 1996). Tiene la ventaja de no requerir supuestos como el de normalidad multivariable y el de homocedasticidad (igualdad de las varianzas), que son difíciles de verificar. Además, es más potente que el análisis discriminante cuando estos supuestos no se cumplen. Otra ventaja radica en su similitud con la regresión múltiple: permite el uso de variables independientes continuas y categóricas (estas últimas por medio de su codificación a variables ficticias), cuenta con contrastes estadísticos directos, tiene capacidad de incorporar efectos no lineales y es útil para realizar diagnósticos (Hair, Anderson, Tatham y Black, 1999). Tiene una amplia aplicación en estudios observacionales, de encuesta y experimentales, como así también en estudios epidemiológicos (Hair et al, 1999; Schelesslman, 1982; Ato y López, 1996; García; Alvarado y Jiménez; 2000). Numerosas investigaciones muestran las ventajas de utilizar el análisis de regresión logística en la evaluación del Funcionamiento Diferencial del Ítem (DIF), especialmente en la detección de DIF cuando es no uniforme y mixto y cuando no se cuenta con muestras grandes (Cortada de Cohan, 2004; Padilla, Gómez e Hidalgo, 2005; Ferreres, Fidalgo y Muñiz, 2000; Hidalgo y López, 1997).

Por lo tanto, teniendo en cuenta el aumento de aplicación de la regresión logística en la investigación psicológica, este trabajo esta orientado a lograr un acercamiento (más práctico que teórico) a los aspectos más importantes relativos a esta técnica estadística, concentrándose en dos aspectos fundamentales, (1) una breve revisión teórica de la técnica, y (2) un ejemplo del uso de la regresión logística, prestando atención a los aspectos práctico que orientan a un uso más adecuado y a una interpretación más confiable de los resultados.

2. El modelo de regresión logística

La regresión logística, al igual que otras técnicas estadísticas multivariadas, da la posibilidad de evaluar la influencia de cada una de las variables independientes sobre la variable dependiente o de respuesta y controlar el efecto del resto. Tendremos, por tanto, una variable dependiente, llamémosla Y, que puede ser dicotómica o politómica y una o más

variables independientes, llamémoslas X, que pueden ser de cualquier naturaleza, cualitativas o cuantitativas. Si la variable Y es dicotómica, podrá tomar el valor "0" si el hecho no ocurre y "1" si el hecho ocurre. Este proceso es denominado *binomial* ya que solo sólo tiene dos posibles resultados, siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones.

Un proceso binomial está caracterizado por la probabilidad de éxito, representada por *p*, la probabilidad de fracaso se representa por *q*. En ocasiones, se usa el cociente *p*/*q* que indica cuánto más probable es el éxito que el fracaso, como parámetro característico de la distribución binomial Los modelos de regresión logística son modelos de regresión que permiten estudiar si una variable categórica depende, o no, de otra u otras variables. La distribución condicional de la variable dependiente, al ser categórica, no puede distribuirse normalmente, toma la forma de una distribución binomial y, en consecuencia la varianza no es constante, encontrándose situaciones de heterocedasticidad. El modelo de regresión logística puede ser representado de la siguiente manera:

$$\text{logist}(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

Donde: $\pi_{1,}$ es la probabilidad de observar la categoría o evento a predecir, y 1-π, es la probabilidad de no observar la categoría o evento a predecir. Es un modelo logístico lineal porque es lineal la escala del logaritmo de la razón de los productos cruzados (RPC). Varía entre $-\infty$ y $+\infty$. Una ventaja de este modelo es que puede utilizarse en muestreos prospectivos o retrospectivos debido a que los efectos se refieren a la razón de los productos cruzados. Para una mejor compresión es conveniente recordar algunos conceptos.

Para ello, tomemos un ejemplo: supongamos que estamos estudiando la malnutrición en niños de 0 a 6 años y su relación con las pautas de crianza. Las categorías de nuestra variable dependiente (Y) eutrófico (que tomará el valor de 0) y mal nutrido (que tomará el valor 1) y para la variable independiente (X) las categorías serán democrático e indiferente (para simplificar el ejemplo no consideraremos la pauta autoritaria). En la tabla 1 se presentan resultados hipotéticos de la evaluación de 400 niños.

Tabla 1.

Evaluación de los niños según estado nutricional y pauta de crianza (N = 400)

Pauta de Crianza	Estado Nutricional		Total
	Malnutrido	Eutrófico	
Indiferente	79 (.395)	121 (.605)	200 (1.00)
Democrático	45 (.225)	155 (.775)	200 (1.00)
Total	106	194	400

La expresión: $\pi_i / (1- \pi_i)$ es la razón de probabilidades (RP) comúnmente denominada *odds*, es la razón entre probabilidades de la variable dependiente para cada uno de los valores de la variable independiente. En nuestro ejemplo la RP (*odds*), para la categoría mal nutrido cuando la pauta de crianza es indiferente es: .395/.605 = .653 y cuando la pauta de crianza es democrática es: .225/.775 = .290. Esto indica que es mayor la probabilidad de observar un niño mal nutrido cuando la pauta de crianza de los padres es negligente.

Un valor de 1 en el *odds* quiere decir que existe equiprobabilidad en ambas categorías de la variable. Un valor mayor que 1 indica que esa categoría tiene mayor probabilidad de ocurrencia. Tiene el inconveniente que su valor varía entre 0 y + ∞. Una transformación logarítmica de la RP u *odds* proporciona una importante medida para el análisis de datos categóricos, denominada transformación logit que varía entre - ∞ y + ∞ y se define sobre una categoría de la variable dependiente (en este caso la categoría esperada, mal nutrido). Así,

$$\text{Logist}\,(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

Debe señalarse que la expresión log se refiere al logaritmo neperiano o logaritmo de base 2 y no al logaritmo decimal. En nuestro ejemplo, para la categoría mal nutrido, en la pauta negligente el logit sería: log (.653) = -.426 .Para la pauta democrática el log (.290) = -1.234. Como el punto central de la escala es 0, es importante considerar el signo para su interpretación.

Más fácilmente interpretable es definir un cociente entre ambas razones de probabilidad (*odds ratio*), que fácilmente se convierte en la razón de productos cruzados (RPC), en nuestro caso (79 x 155) / (45 x 121) = 2.24. Si las probabilidades de observación de mal nutridos en ambos grupos de niños fueran iguales la RPC sería igual a 1, si la razón de

RPC > 1 quiere decir que la probabilidad de observar un mal nutrido es mayor en el primer grupo (pauta paterna negligente) que en el segundo. Un inconveniente es que la RPC no es aditivamente simétrica. Una mayor presencia de mal nutridos para el primer grupo que para el segundo producirá valores de RPC que varían entre 1 y + ∞ mientras que una mayor presencia de mal nutridos en el segundo grupo que en el primero producirá valores de RPC que varían entre 0 y 1, por lo que los rangos no son comparables. Como la RPC es multiplicativamente simétrica, se utiliza el logaritmo natural de la RPC que es una cantidad que varía -∞ y + ∞, siendo 0 donde el efecto es nulo. En nuestro ejemplo el log (2.24) = .806

3. Estimación del modelo

Como se señalara en párrafos anteriores, la distribución condicional de la variable dependiente, al ser categórica, no puede distribuirse normalmente, toma la forma de una distribución binomial. El modelo logístico tiene una forma de curva. Para estimar el modelo se busca la curva que mejor se ajusta a los datos reales. En el gráfico 1 y 2 se muestran dos diagramas de dispersión con diferentes curvas de ajuste. Allí se puede observar cuando la relación es perfecta y cuando la relación es pobre.

Grafico 1.

Relación bien definida

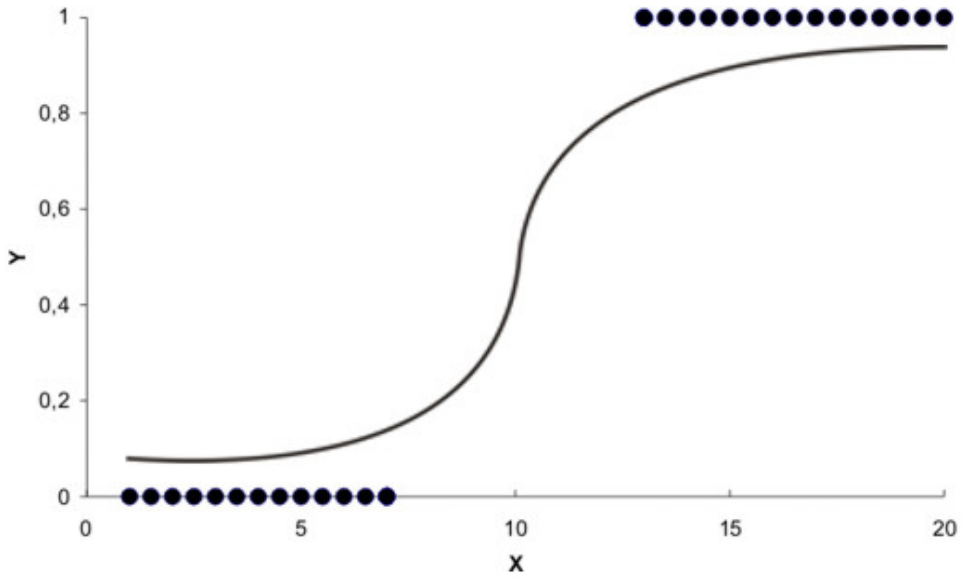
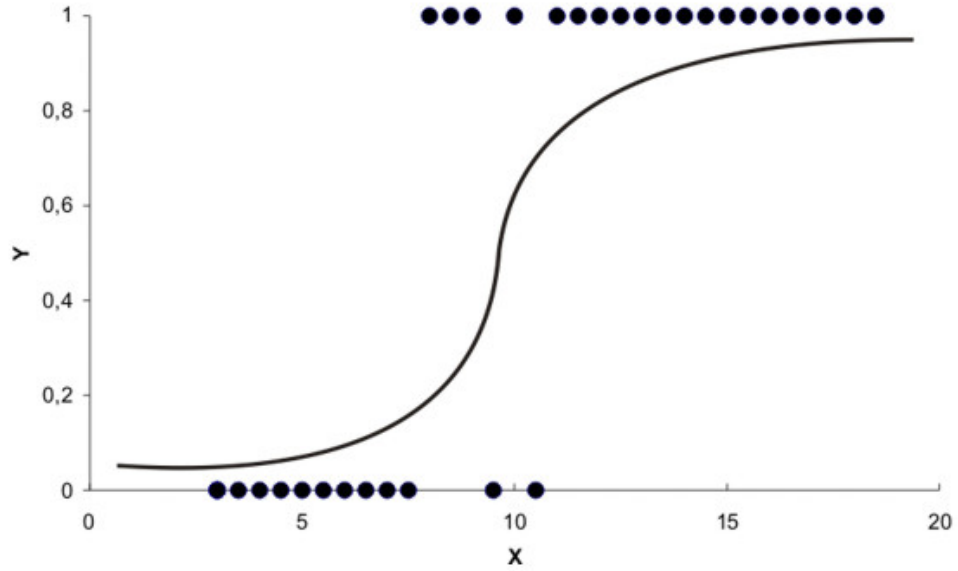


Grafico 2.

Relación pobremente ajustada



Para la estimación del modelo se emplea el método de estimación por máxima verosimilitud que no establece restricción alguna respecto de las características de las variables predictoras, éstas pueden ser nominales, ordinales o intervalares (recuérdese que en la regresión múltiple se utiliza el procedimiento de los mínimos cuadráticos). En el procedimiento de máxima verosimilitud se seleccionan las estimaciones de los parámetros que hagan posible que los resultados observados sean lo más verosímiles posibles. A la probabilidad de los resultados observados, dadas las estimaciones de los parámetros, se la denomina verosimilitud. Como la verosimilitud es un valor pequeño se utiliza como medida de ajuste del modelo a los datos "-2 veces el logaritmo de la verosimilitud" o –2LL. Un buen modelo es aquel que da lugar a una verosimilitud grande por lo cual el valor de –2LL será pequeño.

Se utiliza Chi cuadrado para contrastar la reducción en el valor cuando se introduce una variable independiente. Se compara la diferencia entre (-2LL) o desviancia del modelo inicial (denominado nulo) sin la inclusión de variable predictora alguna y la desviancia del modelo al incluir una o más variables predictoras. Chi cuadrado contrasta la hipótesis nula, postula que los coeficientes de todos los términos excepto la constante son cero. Los grados de libertad en este caso están dados por la diferencia entre el número de los parámetros de los dos modelos.

Hosmer y Lebeshow (1989) han desarrollado una prueba de la bondad del ajuste en relación a la clasificación. El procedimiento consiste en dividir los casos en aproximadamente 10 clases y comparar para cada clase las frecuencias de los casos observados con los casos predichos, utilizando para ello Chi cuadrado. Este procedimiento proporciona una medida global de la capacidad predictiva del modelo que no se basa en el valor de verosimilitud sino en la predicción real de la variable dependiente. Una restricción en su uso es que se necesita contar con una muestra grande que asegure por lo menos cinco observaciones en cada grupo. Por otro lado Chi cuadrado es sensible al tamaño muestral y se puede encontrar significación estadística en diferencias pequeñas al aumentar el tamaño de la muestra. Algunos autores como Hair y Anderson (1999) recomiendan utilizar varios procedimientos para evaluar la bondad del ajuste del modelo.

Para evaluar el ajuste global se han construido medidas similares al coeficiente de determinación (Hair y colaboradores, 1999; Ato y López, 1996), en donde se define al coeficiente de determinación de la siguiente manera:

$$R^2_L = \frac{-2LL_{(\text{nulo})} - 2LL_{(\text{modelo})}}{-2LL_{(\text{nulo})}}$$

Donde: $-2LL_{(\text{nulo})}$: es 2 veces el logaritmo de la verosimilitud del modelo nulo o inicial y $-2LL_{(\text{modelo})}$: es 2 veces el logaritmo de la verosimilitud del modelo a evaluar. El valor de $-2LL_{(\text{nulo})}$ es equivalente a la Suma de los Cuadrados Total en la regresión lineal y el valor de $-2LL_{(\text{modelo})}$ es equivalente a la Suma de los Cuadrados Residual. Este coeficiente es una medida aproximada de la eficacia predictiva del modelo. Como un coeficiente de determinación, cuando la explicación de la varianza de la variable dependiente por el predictor es nula el $R^2_L = 0$ y cuando es perfecta $R^2_L = 1$. Sin embargo, hay que ser cuidadosos en la interpretación porque la variación en el coeficiente de la regresión logística es diferente. Ato y López (1996) señalan que el ajuste lineal suele producir un coeficiente de determinación mayor, por lo cual el coeficiente R^2_L subestima la proporción de varianza explicada por el modelo de regresión logística. En paquetes estadísticos como el SPSS se presentan dos modificaciones de este coeficiente. Uno de ellos es el estadístico Coeficiente R^2_L de Cox y Snell que se computa de la siguiente manera:

$$R^2_L = 1 - \left[\frac{-2LL_{(\text{nulo})}}{-2LL_{(\text{modelo})}}\right]^{2/N}$$

Donde: -2LL $_{(nulo)}$ es la desviancia del modelo nulo solo o con una constante, sin incorporar las variables predictoras, –2LL $_{(modelo)}$ es la desviancia del modelo con las variables predictoras y N es el tamaño de la muestra. Como el valor máximo de esta medida no alcanza 1, Nagelkerke propuso una modificación que incrementa el coeficiente de Cox y Snell para obtener un valor máximo de 1.

Coeficiente $\overline{R}^2_L$ de Nagelkerke

$$\overline{R}^2_L = \frac{1-\left[\frac{-2LL_{(nulo)}}{-2LL_{(modelo)}}\right]^{2/N}}{1-\left(2LL_{((modelo)}\right)^{2/N}}$$

4. El valor teórico y la interpretación de los coeficientes

Como en todo análisis multivariante, en la regresión logística se obtiene un valor teórico, o una combinación lineal de variables con ponderaciones determinadas empíricamente (Hair et al, 1999). La forma del valor teórico de la regresión logística es similar al de la regresión múltiple, y representa una única relación multivariante con coeficientes que indican el peso relativo que tiene cada variable predictora.

$$\text{Logist}\,(\pi_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_i X_j$$

Donde: $\pi_{i:}$ es la proporción (probabilidad de observar la categoría o evento a predecir).

β_0: es una constante

β_1, β_2...β_j: son los coeficientes logísticos correspondientes a cada variable predoctora

X_1, X_2...X_j son las variables predictoras.

La ecuación puede ser presentada en su forma aditiva:

$$\text{Log}\,(\pi_i/1\text{-}\pi_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j$$

También puede ser presentada en su forma multiplicativa:

$$\pi / 1 - \pi = e^{\beta_0} * e^{\beta_1 X_1} * .... * e^{\beta_i X_j}$$

Como se señalara al inicio, en la regresión logística la variable dependiente es dicotómica, por lo cual lo que interesa es predecir la probabilidad de ocurrencia de un evento. El procedimiento de cálculo del coeficiente logístico compara la probabilidad de ocurrencia de un suceso con la probalidad de que no ocurra. Los coeficientes B son las medidas de los cambios en la razón de probabilidad denominado "odds ratio".Están expresados en logaritmos y deben ser transformados para ser interpretados. Un coeficiente positivo aumenta la probabilidad de ocurrencia, y un coeficiente negativo la disminuye. Para la contrastación

de los coeficientes se utiliza el estadístico W de Wald, que es igual al cuadrado de la razón entre un coeficiente de regresión y su error típico. El estadístico W sigue una distribución Chi cuadrado, con un grado de libertad, lo que es apropiado para su uso con datos categóricos

5. Un ejemplo de regresión logística

A continuación se muestra un ejemplo del uso de la regresión logística en una investigación sobre las características del lenguaje en niños con autismo y niños con trastornos en el desarrollo del lenguaje (Gonzalez, 2004). No es nuestro objetivo discutir sus resultados desde el punto de vista clínico, solo pretendemos lograr un mayor acercamiento a los aspectos prácticos anteriormente planteados. Para este estudio se utilizó la regresión logística para identificar las variables que mejor predecían el tipo de trastorno. La variable dependiente *tipo de trastorno* es una variable categórica dicotómica: autismo-disfasia. En función de análisis previos se incluyeron como variables independientes o predictoras, las siguientes variables ordinales continuas: *Compresión, Fonología, Sintaxis, Semántica y Pragmática* (variables ordinales continuas, con las siguientes categorías: Normal: 0, Afectación Leve:1, Afectación Moderada: 2, Afectación Severa: 4) y las variables categóricas dicotómicas: *Ecolalia, Perseveraciones, Inversión Pronominal y Disprosodia*, (categorizadas, Ausencia: 0 y Presencia: 1). Se realizó un trabajo retrospectivo sobre una muestra de 200 niños de ambos sexos A continuación se presentan los resultados obtenidos utilizando el paquete estadístico para las ciencias sociales (Social Package Social Sciences, SPSS). No obstante, este análisis puede ser ejecutado con otros programas tales como SAS o Statistical.

Las primeras dos tablas que ofrece el programa, que no fue incluida en este trabajo, son las que contienen el número de casos incluidos en el análisis y la codificación de la variable dependiente. Es importante tener en cuenta que la categoría a la que le corresponde el código 1 es la categoría sobre la que se calcula la probabilidad de ocurrencia (evento favorable), en este caso es la Disfasia o Trastorno en el desarrollo del lenguaje. A continuación se presentan los datos del modelo nulo es decir los datos sin incluir la información de las variables predictoras. Se realiza la predicción con la única información de los datos observados de la variable dependiente.

Tabla 2.

Tabla de clasificación de los casos observados

Tabla de clasificación[a,b]

Casos Observados			Casos Pronosticados		
			Tipo de Trastorno		Porcentaje Correcto
			Autismo	Disfasia	
Paso 0	Tipo de Trastorno	Autismo	105	0	100,0
		Disfasia	95	0	,0
	Porcentaje Global				52,5

a. Se incluye constante en el modelo

b. El valor del punto de corte es: 0,50

En la tabla 2 se presenta la clasificación de los casos según su ocurrencia y según la predicción realizada en función del modelo nulo. Como puede observarse, habría un 100% de acierto del pronóstico de Autismo y ningún acierto en el pronóstico de la disfasia, por lo cual el porcentaje total de acierto es de 52%.

Tabla 3.

Estadísticos estimados del modelo nulo (Paso 0)

		B	S.E.	Wald	df	Sig.	Exp(B)
Paso 0	Constante	-,100	,142	,500	1	,480	,905

En la tabla 3 se presentan los parámetros del modelo nulo: B o constante, el error estándar correspondiente, el estadístico Wald, los grados de libertad del estadístico, el nivel de significación y el Exponencial de B. El estadístico Wald no es significativo, es decir que B no difiere significativamente de 0 y por lo tanto no produce cambio sobre la variable dependiente. También el programa proporciona datos sobre las variables que no fueron incorporadas en la ecuación. Seguidamente se presentan los datos según que el método de introducción de las variables escogido sea introducir las variables simultáneamente o por pasos. En este caso se utilizó el método por pasos hacia delante utilizando como criterio la significación estadística de los coeficientes B de las variables introducidas usando el estadístico W de Wald.

Tabla 4.

Pruebas ómnibus sobre los coeficientes del Modelo

		Chi-square	df	Sig.
Paso 1	Paso	216,912	1	,000
	Bloque	216,912	1	,000
	Modelo	216,912	1	,000
Paso 2	Paso	8,924	1	,003
	Bloque	225,836	2	,000
	Modelo	225,836	2	,000
Paso 3	Paso	5,043	1	,025
	Bloque	230,879	3	,000
	Modelo	230,879	3	,000

En la tabla 4 se presentan los resultados de las pruebas sobre la disminución de las desviancias (-2LL) o lo que es igual la ganancia obtenida en cada modelo. Recuérdese que cuanto menor es –2LL mejor el ajuste del modelo. Se puede observar para cada paso los valores de las siguientes entradas: Paso, Bloque y Modelo. El Chi cuadrado correspondiente a la fila modelo es la diferencia entre el –2LL para el modelo nulo y –2LL para el modelo actual. Se contrasta la hipótesis nula que postula que los coeficientes de todos los términos excepto la constante son igual a 0 (esto es comparable al test F global para la regresión múltiple). El Chi cuadrado correspondiente a la fila bloque es la diferencia entre 2l-2LL entre los bloques de entrada sucesivos en la construcción del modelo. Como en general se introducen variables en un solo bloque el Chi cuadrado del modelo coincide con el Chi cuadrado del bloque. En la fila correspondiente a Paso el Chi cuadrado es la diferencia entre el –2LL entre pasos sucesivos. Se somete a prueba la hipótesis que los coeficientes de las variables introducidas en el último paso son igual a 0 (comparable al F de cambio en la regresión múltiple). En el ejemplo en consideración en todos los caso el Chi cuadrado es significativo desechándose las hipótesis contrastadas. En la tabla 5 se puede observar resumen de los modelos.

Tabla 5.

Resumen de los modelos

Paso	-2 Log de la verosimilitud	Cox & Snell R Cuadrado	Nagelkerke R Cuadrado
1	59,846	,662	,883
2	50,922	,677	,903
3	45,879	,685	,914

Obsérvese que hay una disminución del –2LL, en relación al primer paso, el -2LL menor corresponde al tercer modelo. Recuérdese que cuando menor es -2 LL, mayor es la verosimilitud y mejor el ajuste del modelo. Los coeficientes de determinación R^2_L son altos y aumentan en el segundo y tercer modelo. El coeficiente de Nagelkerke del último modelo es .914, teniendo en cuenta que el valor máximo es 1 podemos decir que es un coeficiente alto, que un importante porcentaje de la varianza es explicada por las variables predictoras introducidas en el modelo. El programa proporciona también los resultados de la prueba de Hosmer y Lemeshow para cada modelo, basado en la comparación entre los casos observados y los casos pronosticados.

Tabla 6.

Prueba de Hosmer y Lemeshow

Paso	Chi-square	df	Sig.
1	16,028	2	,000
2	3,912	4	,418
3	2,196	5	,821

Como se observa en la tabla 6 para el primer modelo Chi cuadrado es significativo lo cual estaría indicando un mal ajuste del modelo, en el sentido que la hipótesis que se contrasta es que no existen diferencias entre las frecuencias de los casos observados y las frecuencias de los casos pronosticados. En los otros modelos, y en mayor medida en el último, la diferencia no es significativa.

A fin de analizar el ajuste en la clasificación en cada modelo el programa suministra una tabla de clasificación (ver tabla 7) donde se consignan las frecuencias en las categorías de la variable dependiente según lo observado y según lo pronosticado en cada modelo. Los datos proporcionados datos permiten también analizar la especificidad y sensibilidad del modelo y también las tasas de falsos positivos y falsos negativos. En este caso si bien en el modelo 1 y el modelo 3 el porcentaje global es el mismo: 94,5%, en el modelo 3 los falsos positivos y los falsos negativos son menores. Es de destacar que aumenta significativamente el porcentaje global de clasificación correcta en comparación con el porcentaje del modelo nulo que era de 52 %.

Tabla 7.

Tabla de Clasificación

Classification Table[a]

			Casos pronosticados		
			Tipo de Trastorno		Porcentaje Correcto
	Casos Observados		Autismo	Disfasia	
Paso 1	Tipo de Trastorno	Autismo	105	0	100,0
		Disfasia	11	84	88,4
	Porcentaje Global				94,5
Paso 2	Tipo de Trastorno	Autismo	94	11	89,5
		Disfasia	3	92	96,8
	Porcentaje Global				93,0
Paso 3	Tipo de Trastorno	Autismo	98	7	93,3
		Disfasia	4	91	95,8
	Porcentaje Global				94,5

a. Valor del punto de corte 0,50

En una tabla 8 denominada "Variables en la Ecuación" se presentan los estimadores de los parámetros (coeficientes B), sus errores típicos, el estadístico W de Wald, sus grados de libertad y su probabilidad asociada, las estimaciones de las *odds* ratio (Exp B) para las variables predictoras y la constante para cada modelo.

Tabla 8.

Variables en la Ecuación

		B	S.E.	Wald	df	Sig.	Exp(B)
Paso 1[a]	Pragmática	-5,345	1,048	25,988	1	,000	,005
	Constante	4,483	1,006	19,864	1	,000	88,481
Paso 2[b]	Pragmática	-4,682	1,026	20,822	1	,000	,009
	Disprosodia (1)	2,263	,858	6,952	1	,008	9,608
	Constante	2,428	1,270	3,656	1	,056	11,332
Paso 3[c]	Pragmática	-4,340	1,011	18,439	1	,000	,013
	Inversión Pronominal (1)	2,187	1,149	3,622	1	,057	8,912
	Disprosodia(1)	2,114	,892	5,617	1	,018	8,277
	Constante	,509	1,677	,092	1	,761	1,664

a. Variable(s) que entra en el primer paso: Pragmática

b. Variable(s) que entra en el segundo paso: Disprosodia

c. Variable(s) que entra en el tercer paso: Inversión Pronominal

La primera variable que ingresa en el modelo es Pragmática, que presenta el mayor puntaje del estadístico. Analizando el último paso o último modelo podemos ver que fueron incorporadas solamente tres variables predictoras: Pragmática, Inversión y Disprosodia, el resto de las variables fueron desechadas porque sus coeficientes no difieren

significativamente de 0, vale decir no aportan a la predicción de los trastornos en estudio. Observando los coeficientes B podemos postular que la variable que más aporta es Pragmática seguida de Inversión y Disprosodia (estas últimas con mínimas diferencias). Con los estimadores estamos en condiciones de expresar la ecuación predictiva en términos de unidades de la escala logit:

Logit (p) = .509 - 4.340 (Pragmática)+ 1,149 (Inversión)+ .892 (Disprosodia)

Estos coeficientes pueden ser interpretados a la manera de los coeficientes de la regresión múltiple: un incremento en una unidad en la escala de medida de la variable pragmática, produce una disminución de 4, 340 unidades logit de la variable dependiente. El logit es, como se señaló anteriormente, el logaritmo de los productos cruzados. Una interpretación más sencilla consiste en definir el coeficiente en términos de un efecto multiplicativo sobre la escala de la razón de los productos cruzados: e^{β_1} o Exp (B), en nuestro ejemplo un incremento de una unidad en la escala de medida de la variable Pragmática produce un incremento mutiplicativo por un factor de .013 en la escala de los productos cruzados (hay que recordar que esta escala varía entre 0 y + ∞ y que un factor mayor que 1 produce un incremento y que un factor menor que 1 una disminución). En este caso no debe olvidarse que la variable dependiente es Tipo de trastorno y que la categoría con valor 1 es Disfasia o Trastorno en el Desarrollo del Lenguaje. Esto quiere decir que a medida que aumenta la afectación en el aspecto pragmático del lenguaje disminuye la probabilidad de pronóstico de disfasia y, por consiguiente aumenta el pronóstico de autismo. Si esto se traduce en términos de porcentaje de cambio, la interpretación es más accesible, esto es:

$100^* \ (e^{\beta_1} - 1)$ o 100* (Exp (B) – 1) = 100 * ($e^{-4,340} - 1$) = 100* (.013 –1) = - 98,7 %.

En este caso, al aumentar en una unidad la variable predictora Pragmática el porcentaje de cambio (disminución por ser negativo) es del 98,7%. Según los datos obtenidos estamos en condiciones de estimar la probabilidad de ser difásico, conociendo que tiene afectación leve en la expresión pragmática, no presenta Inversión y presenta Disprosodia:

$$\text{P disf/pragmática}_{(1)}, \text{Inversión}_{(0)}, \text{Disprosodia}_{(1)} = \frac{e^{.509-4.340(1)+2.187(0)+2.114(1)}}{1+e^{.509-4.340(1)+2,187(0)+2.114(1)}} = \frac{e^{-1.717}}{1+e^{-1.717}} = \frac{.1796}{1.1796} = .15$$

No debe olvidarse que la variable dependiente es dicotómica y que el punto de corte es p =.50, por lo cual valores iguales o mayores de .50 llevan a pronosticar “Disfasia” codificada 1, en la variable tipo de trastorno y valores menores de .50 “Autismo”, codificada

como 0. Del análisis de la tabla de clasificación del último modelo obtenemos la especificidad (proporción entre frecuencia de aciertos negativos y frecuencia total de negativos observados), la sensibilidad (razón entre la frecuencia de aciertos positivos y la frecuencia total de positivos observados) y también la proporción de falsos negativos y falsos positivos. En el ejemplo sería:

Especificidad: 93.3

Sensibilidad: .95.8

Proporción de falso negativo: .03

Proporción de falso positivo: .07

De las datos anteriores podemos señalar que las variables del modelo tienen alta sensibilidad para diagnosticar adecuadamente la disfasia, hay una proporción alta acierto en el diagnóstico (96%), también es alta su especificidad es decir su capacidad de detectar los casos que no son disfasia (en este caso autismo) y tiene muy bajo porcentaje de error disfasia cuando el trastorno es autismo (7%) y mucho menor de cometer el error de diagnosticar autismo cundo el trastorno es disfasia (3%).

6. Discusión

En síntesis se podría afirmar que la regresión logística es una adecuada alternativa a la regresión lineal cuando lo que se desea predecir es el comportamiento de una variable dependiente categórica. También ofrece ventajas frente al análisis discriminante al no requerir los supuestos de normalidad y homocedasticidad. Para la estimación del modelo se utiliza el procedimiento de máxima verosimilitud. El valor teórico presenta coeficientes que informan el aporte de cada variable predictora en el pronóstico de ocurrencia de las categorías de la variable dependiente. Se han desarrollado varios procedimientos para evaluar la bondad del ajuste del modelo y también se cuenta con coeficientes de determinación al estilo de R^2 en la regresión lineal. Tiene una amplia aplicación en la investigación psicológica y especialmente en los estudios psicométricos.

Referencias

Ato García, M. Y López García, J. J. (1996). *Análisis estadístico para datos categóricos.* Madrid. Editorial Síntesis.

Cortada de Cohan, N. (2004). Teoría de Respuesta al Ítem. *Evaluar,* 4, 95-110.

Ferreres Traver, D; Hidalgo Aliste, A. M. y Muñiz, J. (2000). Detección del Funcionamiento

Diferencial de los ítems no uniforme: comparación de los métodos Mantel-Haenszel y regresión logística. *Psicothema,* 12, 2, 220-225.

García Jiménez, M.V.; Alvarado Izquierdo, J. M. y Jiménez Blanco, J. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema,* 12, 2, 248-252

Gonzalez, G. E. (2004). *Características del lenguaje en niños con autismo y en niños con trastornos del desarrollo del lenguaje. Estudio comparativo.* Tesis de Maestría. Universidad Nacional de Córdoba, Argentina.

Hair, J.F.; Anderson, R.E.; Tatham,R.L.; Black W.C. (1999). *Análisis Multivariante.* 5° Edición. Madrid: Prentice Hall.

Hidalgo Montesinos, M. D. y López Pina, J. A. (1997). Comparación entre las medias de área, el estadístico de Lord y el análisis de regresión logística en la evaluación del funcionamiento diferencial de los ítems. *Psicothema,* 9, 2, 417-431.

Hosmer, D. W., y Lebeshow, S. (1989). *Applied Logistic Regresion.* New York: Wiley.

Pérez, C. (2001). *Técnicas estadísticas con SPSS.* Madrid: Pearson Educación.

Schlesslman, J. J. (1982). *Case–control studies. Desig,Conduct, Análisis.* Nueva York: Oxford University Press.