https://doi.org/10.29393/CE2-3MMJM10003

CIENCIA y ENFERMERIA II (1): 27-40, 1996 I.S.S.N. 0717 - 2079

MODELIZACION MULTIVARIADA DE VARIABLES DEPENDIENTES CATEGORICAS: UNA INTRODUCCION A LA REGRESION LOGISTICA

MULTIVARIATE MODELLING OF CATEGORICAL DEPENDENT VARIABLES: AN INTRODUCTION TO LOGISTIC REGRESSION

JOSE MANUEL MERINO ESCOBAR[1], Ph.D.

RESUMEN

Este artículo se propone presentar los elementos básicos de una de las técnicas más recientes para el análisis de variables dependientes categóricas binarias: el modelo de regresión logística. Utilizando un enfoque relativamente auto-contenido se ha tratado de mostrar los elementos más importantes del modelo sin recurrir a demostraciones algebraicas complicadas.

Dos ilustraciones, relativamente conocidas en la literatura internacional, fueron resumidamente desarrolladas con el propósito de ejemplificar los procedimientos computacionales y de selección y contraste de modelos así también como las perspectivas más generalizadas de interpretación y análisis de los coeficientes estimados en la modelización.

De este modo se divulga en la comunidad científica de las ciencias sociales y de la salud, una estrategia técnica llamada a tener en los próximos años un carácter de instrumento analítico indispensable en el arsenal de los profesionales de estas áreas.

Palabras claves: Regresión Logística. Variables categóricas. Modelización estadística. Razón de verosimilitud.

ABSTRACT

This paper is going to present the basic elements of one of the most recent techniques for analyzing the categorical binary dependent variables: the logistic regression model. By utilizing a relatively self-contained approach was tried to show the models more important dimensions with no reference to complex algebraic demonstrations. Two models illustrations were developed, in summary terms, to show the computational procedures, model contrast and selection, and also the most generalized perspectives for parameters interpretation and analysis which are implied in the modeling process.

In this way is presented to the local scientific community of health and social sciences a technical strategy which in the next years will have a considerable development becoming in a very important analytical instrument in the toolkit of these professionals.

Keywords: Logistic Regression. Categorical variables. Statistical Modeling. Likelihood ratio.

[1]Philosophical Doctor (Ph.D.) in Sociology with a major in Quantitative Methodology. The University of Texas at Austin. Profesor Asociado de la Facultad de Ciencias Sociales de la Universidad de Concepción.

INTRODUCCION

Probablemente el área de mayor desarrollo en la metodología de la investigación científica, en los últimos 20 años, está focalizada en el área del análisis de los datos. Sin embargo, no todas las metodologías de análisis han tenido el crecimiento espectacular que ha caracterizado al análisis de datos categóricos. (Agresti, 1990; Lindsey, 1995). En años recientes la aparición de métodos estadísticos nuevos para el análisis de datos categóricos se incrementó dramáticamente, particularmente de aquellos con aplicaciones directas en las ciencias sociales y biomédicas. Ciertamente esto tiene mucho que ver con el vertiginoso avance de la computación, puesto que muchas de las operaciones implicadas en el manejo de variables categóricas son prácticamente imposibles de ejecutar sin procesamiento computacional.

Conceptos tales como regresión logística, análisis loglineal, logit, razón de odds, razón de verosimilitud, variables limitadas, regresión Poisson, etc., son términos que cada día adquieren más importancia en los últimos números de las revistas de frontera de las ciencias sociales y de la salud. Todos ellos están referidos al análisis de datos categóricos y deberían ser conocidos y utilizados por quienes están interesados en la práctica de la investigación científica en los campos antes enunciados.

El presente artículo pretende simplemente hacer una breve introducción al análisis actual de datos categóricos con el propósito de extender su uso entre los profesionales de las ciencias de la salud quienes deben frecuentemente utilizar este tipo de variables en su quehacer cotidiano. Esta presentación sin embargo pretende a la vez demostrar la importancia de la modelización multivariada de datos categóricos como la estrategia de análisis más relevante para muchos fenómenos sociales y de salud. Probablemente, este segundo objetivo debería ser especialmente observado por quienes tienen responsabilidad de ejecutar y/o evaluar investigaciones al nivel graduado, y están por tanto sujetos a mayores exigencias de calidad y precisión.

En particular, en este artículo se hará una rápida introducción a la regresión logística que constituye hoy una de las herramientas más avanzadas en la modelización de variables categóricas (Hosmer & Lemeshow, 1989). La idea central será exponer sintéticamente sus principales características tanto como ejemplificar su utilidad empírica mediante un ejercicio de modelización que demuestre el procesamiento computacional necesario para hacer las estimaciones estadísticas e ilustre adecuadamente acerca de los modos en que deben interpretarse los coeficientes obtenidos (McDermott & Whyte, 1994).

DATOS CATEGORICOS

Comencemos por definir qué debe ser entendido por datos categóricos. Una variable categórica es aquella cuya escala de medición consiste básicamente de un conjunto de categorías. Por ejemplo, un test diagnóstico para la enfermedad de Alzheimer podría utilizar como categorías clasificatorias "síntomas presentes" y "síntomas ausentes"; la filosofía política de una persona podría ser clasificada como "liberal", "moderada" o "conservadora"; y la elección del tipo de desayuno podría ser entre "frío", "caliente" o "ninguno". Todas estas clasificaciones son ejemplos de categorías "mutuamente excluyentes" (cada individuo puede ser adscrito a una y sólo a una de estas alternativas) y "exhaustivas" (todos los individuos deben ser clasificados en alguna de las alternativas), que constituyen las principales características de los principios clasificatorios que debe reunir la medición categórica (Agresti, 1996).

Las escalas categóricas son ampliamente utilizadas en las ciencias sociales para medir actitudes y opiniones de todos los tipos. También estas escalas son muy frecuentemente usadas en las ciencias de la salud para medir respuestas tales como si un paciente sobrevive a una operación quirúrgica (sí, no); la se-

veridad de una lesión (ninguna, leve, moderada, severa); o el ciclo o etapa de una enfermedad (inicial, avanzado). Aunque son más frecuentes en las ciencias sociales y de la salud, estas variables no estan restringidas sólo a estas disciplinas. Ellas son también usadas muy frecuentemente en psiquiatría (por ejemplo, en el uso de las categorías "esquizofrenia", "depresión", "neurosis" para la conceptualización del tipo de enfermedades mentales), salud pública (como cuando se usan las categorías "sí", "no" para indicar que el conocimiento del SIDA ha influido en el uso de condones), zoología (uso de categorías "peces", "invertebrados", "reptiles" para expresar las preferencias de alimentación de los cocodrilos), educación (por ejemplo, uso de las categorías "correcto" e "incorrecto" para las respuestas de un estudiante en un examen), marketing (utilización de categorías "marca A" , "marca B", "marca C" para medir la preferencia de los consumidores entre las tres marcas principales de un producto).

La modelización estadística utilizada hoy en la investigación científica divide las variables entre las consideradas como dependientes o respuestas y aquellas consideradas predictoras, explicativas o independientes (Dobson, 1990). Entre las dos conceptualizaciones propuestas es más relevante hablar de variables respuestas y variables predictoras. El propósito de la modelización estadística es demostrar cómo la variación en los valores observados de la variable respuesta (dependiente) puede ser explicada por diferencias en circunstancias entre individuos o grupos de individuos. Estas diferencias pueden deberse a la naturaleza (sexo, edad), ser creadas por el hombre (medio ambiente social, educación), debidas a exposición (dieta, polución, medios de comunicación masivos), o a tratamientos (medicamentos, fertilizantes, técnicas educativas) (Collet, 1991). Todas estas diferencias son representadas por las variables predictoras. El principal punto aquí es que las variables categóricas juegan un muy importante papel en ambos tipos de conceptualizaciones. Sin embargo, en este artículo a nosotros nos interesa esencialmente demostrar cómo pueden ser modeladas eficientemente variables respuestas categóricas. Esto es, nuestro propósito es introducir el modelo estadístico más apropiado para explicar en términos multivariados una variable dependiente categórica. Con una sola restricción: la variable dependiente o respuesta categórica debe ser binaria (dicotómica). En este caso el modelo estadístico apropiado es la regresión logística multivariada. Cuando la variable respuesta (dependiente) categórica tiene más de dos categorías debe utilizarse la regresión logística multinomial, versión especializada de la regresión logística que amerita un artículo independiente.

REGRESION LOGISTICA

En la investigación corriente de las ciencias sociales y de la salud es muy frecuente encontrar situaciones en que la variable respuesta de interés es una dicotomía tal como vivo o muerto, divorciado o aún en el matrimonio, acepta o rechaza la anticoncepción, etc. En los últimos tiempos, la regresión logística ha sido utilizada para estudiar tópicos tan diversos como la formación y disolución marital (Abdelrahman & Morgan, 1987; White, 1987; Trussel & Rao, 1989), uso de anticonceptivos (Studer and Thorton, 1987; Bean, *et. al.*, 1987; Merino, 1993), miseria (Smith & Zick, 1986); experiencias sexuales pre-matrimoniales (Newcomber & Udry, 1987), embarazo pre-matrimonial (Robbins et al., 1985; Yamaguchi & Kandel, 1987; McLanahan, 1988), matrimonios sin hijos (Rao, 1987), abuso en el matrimonio (Kalmuss & Seltzer, 1986).

Todos estos estudios que han empleado regresión logística han tenido en común la variable dependiente o respuesta binaria. Las variables predictoras en cambio han sido cuantitativas, categóricas o una mezcla de ambos tipos. Esta técnica es análoga a la regresión lineal en que una variable respuesta continua es modelada como una función li-

neal de un conjunto de predictores continuos (Liao, 1994). Existen, sin embargo muy importantes diferencias que las hacen específicas. La principal consiste en que la variable dependiente de la regresión logística sólo tiene valores en el intervalo (0,1). Mediante el uso de una variable dummy el 0 ha sido asignado a la categoría "fracaso" (muerto, enfermo, divorciado, etc.) mientras el 1 ha sido utilizado para codificar los casos "éxito" (vivo, sano, casado, etc.).

Si la variable dependiente binaria es un dummy (0,1) es posible entonces ajustar el modelo de regresión lineal. Este uso de la regresión lineal para una dummy variable dicotómica se llama modelo de probabilidad lineal y a sido utilizado con alguna frecuencia para modelar la variable dependiente binaria. Al utilizar la regresión ordinaria (mínimos cuadrados) para modelar la variable binaria con mucha probabilidad los valores predictivos de Y (variable dependiente) estarán fuera del intervalo (0,1) lo que implica que la probabilidad P asumirá valores imposibles. Al extremo izquierdo la variable dependiente obtendrá valores negativos mientras que en el extremo derecho tendrá valores superiores a 1.

Por otra parte, al ajustar una linea recta, se estará asumiendo en la variable dependiente dicotómica una estructura de error normal en circunstancias que una variable dicotómica tiene una distribución binomial. Necesariamente esta distribución no tendrá varianza constante, esto es, será heterocedástica por lo que violará un tercer supuesto de la regresión mínimo cuadrática ordinaria. Ciertamente para analizar una variable respuesta dicotómica la mejor distribución es la logística. Esta distribución tiene una forma de curva sigmoide, esto es, se parece a una S alargada o una S invertida. Las colas de la curva sigmoide se hacen asíntotas antes de lograr $P=0$ ó $P=1$ de tal manera que los valores imposibles de P (menores que 0 y mayores que 1) son evitados (Retherford & Kim, 1993). Existen dos razones según Cox (1970), para preferir en estas circunstancias la distribución logística. Estas son: (1) desde un punto de vista matemático, esta distribución es extremadamente flexible y fácil de usar y (2) tiene una muy significativa y fácil interpretación biológica.

El concepto necesario para entender la regresión logística es el de razón de chances (odds-ratio) (Morgan & Teachman, 1988). La razón de chances es una muy importante medida de asociación. Como su nombre lo expresa claramente es la razón de dos chances. Las chances son a su vez las razones del número de eventos al número de no eventos. Por ejemplo, si nuestra variable dependiente de interés es disolución marital, las chances (odds) son calculadas como el número de disoluciones matrimoniales al número de matrimonios aún intactos. La conversión desde odds a proporciones es muy fácil y se puede expresar como: O (odds) = P/(1-P), donde P es cualquier proporción. La ecuación logística es aquella en que el logaritmo natural de los odds de la variable dependiente binaria es predicha por una función lineal de las variables independientes o predictoras. La razón de odds, el concepto central del modelo de regresión logística, tiene según Fienberg (1977), varias propiedades como medida de asociación:

1. La razón de odds tiene una clara interpretación. Si dos totales son fijos, la razón de odds da el cambio multiplicativo requerido para pasar desde una chance a la próxima. Un odds ratio mayor que 1.0 sugiere una mayor probabilidad de que ocurra el evento (mayor probabilidad al divorcio), mientras que un odds ratio menor que 1 indica una menor probabilidad de ocurrencia del evento.

2. La razón de odds es invariante al intercambio de hileras o columnas.

3. Es también invariante a la multiplicación de las hileras y columnas. Los cambios en tamaño de la muestra no afectan el valor de los odds ratio.

El modelo de regresión logística supone que cada miembro de la población tiene alguna probabilidad subyacente de éxito en una variable independiente dada. Por lo tanto, en la población, cada miembro con un conjunto dado de características tiene una chance p de éxito y 1-p chances de fracaso. Cuando se usan datos de nivel individual, cada miembro tiene ya sea una chance 1 ó 0 de éxito.

Dejemos que p_i sea la probabilidad que la i-ésima persona en la muestra esté en la categoría de interés (éxito) en una variable dependiente binaria y que $(1-p_i)$ sea la probabilidad que la persona esté en la otra categoría (fracaso). Es claro que $p_i/(1-p_i)$ es igual a los odds de estar en la categoría de interés para el i-ésimo individuo. Ahora bien, $\log(p_i/(1-p_i))$, los log-odds o logit de estar en la categoría de interés es una variable continua que puede tomar cualquier valor en el rango $(-\infty, +\infty)$. Establezcamos también que $X_{i1}, X_{i2}, \ldots, X_{ik}$ sea un conjunto de k predictores continuos o categóricos medidos al i-ésimo individuo en la muestra. Entonces el modelo de regresión logística para los log-odds, dado un vector particular de puntajes sobre los k predictores, es:

$$p_i = \frac{e^{\beta o + \beta 1 X}}{1 + e^{\beta o + \beta 1 X}}$$

Una transformación de p_i que es central al estudio de la regresión logística es la transformación logit. Esta transformación es definida, en términos de p_i, de la siguiente manera:

$$\log(p_i/(1-p_i)) = \beta_o + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

y el correspondiente modelo multiplicativo para los odds, es:

$$p_i/(1-p_i) = e^{\beta o} e\beta^{1X1} e\beta^{2X2} e^{\beta Xk}$$

La función $\log (p_i/(1-p_i))$ es llamada la transformación logística (o logit) y todo el modelo es conocido como modelo de regresión logística. Se utilizan logaritmos naturales (base e) en los modelos logísticos. Así como p_i pasa de 0 a 1, el logit crece de $-\infty$ a ∞. La probabilidad $p_i = 0.50$ corresponde a un logit de 0 y los valores de p_i sobre (abajo) 0.50 corresponden a logits positivos (negativos). Si $\beta = 0$ en el modelo de regresión logística, el logit (y por tanto p_i) no cambia cuando X cambia por lo tanto la curva se aplana hacia una línea horizontal. Para β positivas, p_i crece así como X crece. Para β negativas, p_i decrece cuando X crece, en otras palabras, la probabilidad de una respuesta "1" tiende hacia 0 para valores grandes de X. Algunas veces es útil conocer el valor de X en que $p_i = 0.50$. Esto ocurre cuando $\log(p_i/(1-p_i)) = 0 = \beta_o + \beta_i X_i$, esto es, cuando $X = -\beta_o/\beta_i$.

Las estimaciones de los coeficientes β, coeficientes de regresión, son obtenidos mediante el método de máxima verosimilitud (maximum likelihood). Como en la regresión líneal se deben emplear tests globales para el modelo completo, así como test individuales para la significación de los coeficientes estimados para cada variable o nivel de la variable (Halli, 1992). El test global del modelo es la razón de verosimilitud (likelihood ratio), o L_2, que es igual a $-2 \ln(L_o/L_1)$, donde L_1 es el valor de la función de verosimilitud para el modelo de interés, mientras L_0 es el valor de la función de verosimilitud para el modelo más simple que excluye todos los coeficientes excepto el intercepto. Esta prueba estadística tiene distribución chi-cuadrado bajo la hipótesis nula que todos los coeficientes excepto el intercepto son ceros y, por lo tanto, proporciona una prueba para determinar si alguno de los predictores son necesarios para modelar el evento de interés (Liao, 1994). Cuando L_2 es significativo se debe inferir que al menos unos de los coeficientes $\beta_1\beta_2\ldots\beta_k$ es diferente de cero (Hirsch & Riegelman, 1992).

Los predictores individuales se verifican en términos de significatividad mediante el examen de la razón de sus coeficientes estimados a sus errores estandard, los cuales son aproximadamente normales bajo la hipótesis nula que el coeficiente es cero en la pobla-

ción. La interpretación de los coeficientes de regresión logística es completamente fácil. Por ejemplo, β_1 representa el incremento de los log-odds por cada unidad de incremento en X_1, manteniendo todos las otras variables predictoras constantes. O, alternativamente, e^{B1} representa el factor multiplicativo por el cual los odds cambian por cada unidad de incremento en X_1, controlando por todos los otros predictores (Demaris, 1992; Roncek, 1991).

APLICACIONES DE LA REGRESION LOGISTICA

Desarrollaremos, a continuación, dos ilustraciones que aplican regresión logística en dos situaciones diferentes. La primera puede ser considerada un ejemplo de regresión logística bivariada, esto es, la situación más simple en que se puede utilizar este tipo de modelo estadístico para variables categóricas. La segunda es un ejercicio que utiliza regresión logística múltiple para modelar un conjunto de predictores actuando sobre una variable dependiente binaria. Ambos ejercicios han sido seleccionados en la literatura como aplicaciones apropiadas de esta estrategia para los fines pedagógicos que hemos establecido como principal meta de este artículo.

Ilustración 1

El primer data set analizado ha sido extraído desde Agresti (1990, pp.122-123) y corresponde a un pequeño estudio derivado desde el área de la senilidad. Se hizo un examen psiquiátrico a una muestra de 54 ancianos para determinar si existían entre ellos síntomas de senilidad. La información obtenida fue utilizada para construir una variable binaria que se denominó SINT: 1 = síntomas de senilidad presentes, 0 = sin síntomas de senilidad. Se les aplicó además a los ancianos una subescala de la Wechsler Adult Intelligence Scale (WAIS), para medirles el funcionamiento intelectual. Los puntajes de esta escala varían entre 4 y 20 puntos y los valores superiores indican un mejor funcionamiento intelectual. Esta escala fue utilizada como un predictor continuo. Los datos son observables en la siguiente tabla:

Tabla N° 1. Data sobre X = Puntaje WAIS y Y = Senilidad (1 = síntomas presentes).

X	Y	X	Y	X	Y	X	Y	X	Y
9	1	7	1	7	0	17	0	13	0
13	1	5	1	16	0	14	0	13	0
6	1	14	1	9	0	19	0	9	0
8	1	13	0	9	0	9	0	15	0
10	1	16	0	11	0	11	0	10	0
4	1	10	0	13	0	14	0	11	0
14	1	12	0	15	0	10	0	12	0
8	1	11	0	13	0	16	0	4	0
11	1	14	0	10	0	10	0	14	0
7	1	15	0	11	0	16	0	20	0
9	1	18	0	6	0	14	0		

Regresión Logística en SAS

SAS tiene cuatro procedimientos que permiten calcular la regresión logística: PROC LOGISTIC, PROC GENMOD, PROC CATMOD y PROC NLIN. El PROC LOGISTIC es lejos actualmente el más indicado y ventajoso para modelar la regresión logística, por lo tanto será el procedimiento a utilizar en esta discusión (MacDermott & White, 1994).

El siguiente es el programa SAS que permite leer la data del estudio de senilidad, generar tabulaciones cruzadas de las dos variables y realizar la regresión logística de SINT sobre WAIS:

```
data one;
   infile 'senil.dat';
   input wais sint;
   label wais = 'Puntaje en Escala WAIS';
   label sint = 'Sintomas de senilidad';
proc freq data = one;
   tables wais*sint/nocol norow nocum nopercent;

proc logistic descending;
   model sint = wais/risklimits;
```

Al contrario de otros paquetes estadísticos, SAS automáticamente modela el evento que es igual a 0 (sint = 0). Para cambiar a la situación que nos interesa, esto es, modelar la probabilidad que el evento sea igual a 1 (sint = 1), se debe especificar la opción DESCENDING en la frase proc logistic (Schlotzhauer, 1995). Si no se agrega esta opción, el signo del parametro estimado será el opuesto al que normalmente corresponde. La opción RISKLIMITS agregada en la línea del modelo, solicita a SAS que los coeficientes, además de ser calculados en log-odds (logits), sean computados también en términos de odds ratio, con sus respectivos intervalos de confianza. Sin embargo, esta opción de cálculo directo de los odds ratio es sólo válida en la versión 6.09 de SAS.

Tabla N° 2. Tabulación cruzada de puntajes Wais y síntomas de senilidad. Regresión logística de sint sobre Wais. Output de resultados sas.

Table of Wais by Sint

Wais (Wais Score)	Sint (Symptoms of senility)		
Frequency	0	1	Total
4	1	1	2
5	0	1	1
6	1	1	2
7	1	2	3
8	0	2	2
9	4	2	6
10	5	1	6
11	5	1	6
12	2	0	2
13	5	1	6
14	5	2	7
15	3	0	3
16	4	0	4
17	1	0	1
18	1	0	1
19	1	0	1
20	1	0	1
Total	40	14	54

Sigue a continuación el output de resultados SAS para la regresión logística:

The LOGISTIC Procedure

Data Set: WORK.ONE
Response Variable: SINT Symptoms of senility
Response Levels: 2
Number of Observations: 54
Link Function: Logit

Response Profile

Ordered Value	SINT	Count
1	1	14
2	0	40

Simple Statistics for Explanatory Variables

Variable	Mean	Standard Deviation	Minimum	Maximum	Variable Label
WAIS	11.574074	3.709253	4.00000	20.0000	WAIS Score

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	63.806	55.017	.
SC	65.795	58.995	.
-2 LOG L	61.806	51.017	10.789 with 1 DF (p = 0.0010)
Score	.	.	9.795 with 1 DF (p = 0.0017)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Variable Label
INTERCPT	2.4040	1.1918	4.0687	0.0437	.	Intercept
WAIS	-0.3235	0.1140	8.0570	0.0045	-0.661626	WAIS Score

La observación del output de resultados provenientes del proceso SAS, nos entrega la siguiente ecuación de regresión logística:

$$\log (p_i/(1-pi)) = \beta_o + \beta_1 X = 2.404 - 0.324X$$

con $\beta_o = 2.404$ y $\beta_1 = -0.324$. Dado que la estimación de β_1 es negativa, esta muestra

sugiere que la probabilidad de los síntomas de senilidad decrecen a mayores niveles de WAIS. La hipótesis nula H_o: $\beta_1 = 0$ establece que la probabilidad de senilidad es la misma en todos los niveles de WAIS. El error standard estimado de $\beta 1$ es $\sigma b = .114$. Para probar Ho: $\sigma b = 0$, se debe realizar la prueba estadística $Z = \beta 1/\sigma b = -.324/.114 = -2.84$. Este resultado es significativo para $p = .004$ en una prueba de dos colas, por lo que se debe concluir que existe una fuerte evidencia de relación negativa entre la presencia de senilidad y el puntaje en la escala WAIS.

Determinando probabilidades y odds en la regresión logística

Solucionando la ecuación de regresión logística para p_i, podemos expresar el modelo directamente en términos de p_i como sigue:

$$p_i = e^{\beta o + \beta 1X} / (1 + e^{\beta o + \beta 1X})$$

En esta fórmula e elevado a algunas potencia representa el antilogaritmo de ese número cuando se usan logaritmos naturales. En el ejemplo de la senilidad, la proporción predicha de ancianos con síntomas de vejez es:

$$p_i = e^{2.404 - .324X} / (1 + e^{2.404 - .324X})$$

Las personas que puntúan 4 en la escala WAIS (esto es, las personas con los puntajes más bajos de la muestra en la escala) obtienen una proporción predicha o ajustada de:

$$p_i = e^{2.404 - .324(4)} / (1 + e^{2.404 - .324(4)})$$

$$= e^{1.11} / (1 + e^{1.11}) = 3.03/4.03 = .75$$

Para WAIS = 20 (el valor más alto obtenido en la muestra), la proporción ajustada o predicha es sólo .02. Para el valor promedio de WAIS = 11.6, la proporción predicha es .21. La razon $p_i / (1 - p_i)$ que aparece en la transformacion logit son los odds. Por ejemplo, cuando $p_i = .75$, los odds son iguales a .75/.25 = 3.0, lo que significa que una respuesta de '1' es tres veces más probable que una respuesta de '0'. Una importante vía para interpretar los coeficientes β de regresión logística es como un efecto sobre los odds. Específicamente tomando antilogaritmos de ambos lados de la ecuación logística $\log(p_i/(1 - p_i)) = \beta_o + \beta_1 X$, obtenemos:

$$\frac{p_i}{1-p_i} = e^{\beta o + \beta 1X} = e^{\beta o}(e^{\beta 1})^X$$

El lado derecho de esta ecuación tiene la forma exponencial. Esta relación exponencial implica que cada unidad de incremento en X produce un efecto multiplicativo de e^{β} sobre los odds. En el ejemplo de la senilidad, el antilog de β es estimado ser $e^{\beta} = e^{-.324} = .723$. Cuando WAIS = 20, por ejemplo, los odds para la senilidad son estimados ser .723 veces más lo que ellos son cuando WAIS = 19. Cuando WAIS = 20, los odds de senilidad son sólo $(.723)10 = .04$ veces tan alto como cuando WAIS = 10:

$$\frac{p_i}{1-p_i} = e^{2.404 - .324(10)} = .433$$

en tanto cuando WAIS = 20,

$$\frac{p_i}{1-p_i} = e^{2.404 - .324(20)} = .017$$

valor que es sólo un 4% del valor de .433 que se obtiene cuando WAIS = 10.

Verificación de la bondad de ajuste de los modelos

Hasta aquí nos hemos dedicado a mostrar cómo se deben interpretar los resultados de una regresión logística en términos de sus resultados obtenidos. Ahora completaremos la modelización examinando la bondad del ajuste de los modelos mediante la comparación de sus razones de verosimilitud.

La ecuación de regresión logística que hemos examinado incluye sólo una variable predictor, la Escala WAIS. La pregunta que debe ser respondida es si acaso este modelo es significativo en reducir variación de la variable dependiente. Si la respuesta es positiva, entonces la variable predictora, la escala WAIS, tiene efectos sobre la variable dependiente. Por el contrario, si la variable independiente no reduce significativamente variación de la variable respuesta, entonces la escala WAIS no tiene efectos sobre la variable síntomas de senilidad. Para valorar esta comparación de modelos se hace uso de las razones de verosimilitud de los modelos. En este caso tenemos el llamado "modelo nulo", que se ajusta sin incluir ningún predictor, L_0. El modelo nulo es aquel que ajusta sólo el promedio de la variable dependiente, que tiene como término sólo el intercepto y que sirve, fundamentalmente, còmo baseline para efectuar la comparación de modelos ya que su valor de -2 log L es igual a la cantidad total de variación de la variable dependiente.

En el output SAS de páginas anteriores frente a -2 Log L aparece el valor de 61.806 que es la cantidad total de variación de la variable dependiente. Obsérvese en la misma hilera que cuando se ajusta el modelo que agrega al intercepto sólo la variable WAIS, L1, la cantidad de variación de la variable dependiente es reducida a 51.017, este es el log likelihood del modelo que hemos ajustado. Esto implica que al agregar una variable (WAIS) al modelo nulo hemos reducido 10.789 puntos de variación de la variable dependiente, cantidad que al tener una distribución chi-cuadrado, para un grado de libertad (los grados de libertad del modelo corresponden a la diferencia del número de coeficientes a ser estimado en los dos modelos que se comparan) es una reducción estadísticamente significativa al valor p = .001.

Esto implica llegar a la conclusión final que la escala WAIS –como medida de inteligencia de los ancianos– es un predictor absolutamente importante de la senilidad. Es decir, el modelo examinado es uno que incluye una variable con efectos significativos sobre la variable dependiente. La dirección de esos efectos ha sido ya determinada como inversa al examinar los coeficientes de la ecuación en la sección anterior.

Ilustración 2

El segundo dataset, a utilizar como ejercicio en este artículo, ha sido extraido desde Collet (1991) y se refiere a una modelización del cáncer prostático mediante una regresión logística hecha por Brown en 1980. El régimen de tratamiento que debe ser adoptado para pacientes que han sido diagnosticados con cáncer prostático es crucialmente dependiente de si el cáncer se ha difundido o no a los nódulos linfáticos cercanos. Así, una laparotomía se puede hacer sólo para observar el grado de involucramiento de los nódulos. Existen varias variables que son indicativas de compromiso de los nódulos que pueden ser medidas sin cirugía. Brown desarrolló este estudio para observar si una combinación de 5 variables podría ser usada para pronosticar involucramiento nódular del cáncer. Las 6 variables del estudio fueron:

Age: Edad del paciente en años
Acid: Nivel de serum ácido fosfatasa
X-ray: Resultado de un examen de Rayos X (0 = negativo; 1 = positivo)
Size: Tamaño del tumor (0 = pequeño; 1 = grande)
Grade: Avance del tumor (0 = menos serio; 1 = muy serio)
Nodes: Involucramiento nodular (0 = no; 1 = si)

Los valores de cada una de estas variables fueron medidas para 53 pacientes con cáncer prostático que habían ya sido sometidos previamente a una laparotomía. La data base original se encuentra en el Apéndice 1.

El propósito del estudio es construir un modelo que pueda ser usado para predecir los valores de la variable respuesta binaria, in-

volucramiento nodular, sobre las bases de las cinco variables predictoras antes descritas. Como son 5 las variables predictoras es posible construir 32 modelos alternativos: 1 sin ninguna variable predictora (modelo nulo), 5 modelos con sólo una variable independiente, 10 modelos con dos variables diferentes, 10 variables con tres variables diferentes a la vez, 5 modelos con cuatro variables distintas a la vez, 1 modelo con las cinco variables diferentes simultáneamente.

El siguiente es un extracto del programa SAS que permite modelar la ecuación de regresión logística para determinar la razón de verosimilitud (likelihood ratio) y los coeficientes de cada modelo. Dada la extensión del programa sólo hemos incluido un modelo de cada clase. Obviamente los restantes modelos son simples repeticiones de las clases ejemplificadas:

Programa SAS para hacer regresión logística

```
options linesize = 80;

libname brown '/user1/efecon/jmerino/
discrete';

data brown.pte;
infile '/user1/efecon/jmerino/discrete/
prost.dat';
input id age acid xray size grade nodal;
lacd = log(acid);

proc freq data = brown.pte;
table age*lacd*xray*grade*size*nodal /
list out = tridim;


proc sort data = tridim;
  by descending nodal;

/* Ajustando el modelo nulo */

proc logistic data = tridim
    order = data;
    model nodal = ;
    weight count;

/* Ejemplo de ajuste de modelo con un
efecto independiente */

proc logistic data = tridim
    order = data;
    model nodal = age;
    weight count;

/* Ejemplo de ajuste de modelo de efectos
aditivos con dos variables */

proc logistic data = tridim
    order = data;
    model nodal = age lacd;
    weight count;

/* Ejemplo de ajuste de modelo de efectos
aditivos con tres variables */

proc logistic data = tridim
    order = data;
    model nodal = age lacd xray;
    weight count;

/* Ejemplo de ajuste de modelo de efectos
aditivos con cuatro variables */

proc logistic data = tridim
    order = data;
    model nodal = age lacd xray grade;
    weight count;

/* Ajuste del modelo de efectos aditivos
con cinco variables */

proc logistic data = tridim
    order = data;
    model nodal = age lacd xray grade size;
    weight count;
```

Al correr este programa SAS se obtiene el output de resultados que incluye la deviance, o sea, la razón de verosimilitud, los grados de libertad y los coeficientes pertinentes a cada uno de los modelos. En este caso lo interesante es seleccionar el modelo que mejor ajusta los datos, esto es, aquel que es más eficiente en reducir variación de la variable de-

pendiente introduciendo la menor cantidad de complejidad posible. La tabla siguiente incluye la deviance y los grados de libertad de cada uno de los 32 modelos de regresión logística posibles para estos datos:

Tabla Nº 3. Resultados del ajuste de 32 modelos de regresión logística a los datos sobre involucramiento nodular en pacientes con cáncer prostático.

Términos ajustados en el modelo	Deviance	gl
Constante	70.25	52
Age	69.16	51
log(acid)	64.81	51
X-ray	59.00	51
Size	62.55	51
Grade	66.20	51
Age + log(acid)	63.65	50
Age + X-ray	57.66	50
Age + Size	61.43	50
Age + Grade	65.24	50
log(acid) + X-ray	55.27	50
log(acid) + Size	56.48	50
log(acid) + Grade	59.55	50
X-ray + Size	53.35	50
X-ray + Grade	56.70	50
Size + Grade	61.30	50
Age + log(acid) + X-ray	53.78	49
Age + log(acid) + Size	52.22	49
Age + log(acid) + Grade	58.52	49
Age + X-ray + Size	52.08	49
Age + X-ray + Grade	55.49	49
Age + Size + Grade	60.28	49
log(acid) + X-ray + Size	48.99	49
log(acid) + X-ray + Grade	55.03	49
log(acid) + Size + Grade	54.51	49
X-ray + Size + Grade	52.78	49
Age + log(acid) + X-ray + Size	47.68	48
Age + log(acid) + X-ray + Grade	50.79	48
Age + log(acid) + Size + Grade	53.38	48
Age + X-ray + Size + Grade	51.57	48
log(acid) + X-ray + Size + Grade	47.78	48
Age + log(acid) + X-ray + Size + Grade	46.56	47

La variable predictora individual que reduce más variación de la variable dependiente es X-Ray. Si comparamos la deviance del modelo nulo (70.25) con la deviance del modelo que incluye sólo la variable X-Ray (59.00), encontramos que al agregar esta variable al modelo nulo se reduce en 11.25 puntos la variación de la variable dependiente perdiendo sólo un grado de libertad. Este valor de chi-cuadrado es altamente significativo. El modelo de dos variables que reduce más variación de la variable dependiente es X-Ray + Size con una deviance igual a 53.35. Si comparamos este modelo con el que incluye sólo a X-ray, se puede inferir que Size reduce independientemente otros 5.65 puntos de deviance, lo que también es un valor de chi-cuadrado significativo al 5%, para un grado de libertad.

Entre los modelos con tres variables, el que reduce más variación es aquel que incluye a X-ray + Size + log(acid). Este modelo tiene una deviance de 48.99 puntos. Al compararlo con el anterior modelo de dos variables incluyendo a X-ray + Size, encontramos que este modelo, en verdad, prueba el efecto aditivo de la variable log(acid). Al introducir log(acid) en el modelo de dos variables, la reducción de deviance es igual a 53.35-48.99 = 4.36 puntos. Para una pérdida de un grado de libertad, estos 4.36 puntos son un valor de chi-cuadrado significativo al 5%. Por tanto, la variable log(acid) también contribuye independientemente a reducir la variación de la variable respuesta. Por lo tanto, el modelo con las tres variables enunciadas es significativo en reunir tres variables con efectos independientes sobre el involucramiento nodal.

Entre los modelos con cuatro variables el que reduce más variacion es Age + log(acid) + X-ray + Size, que lleva la deviance a 47.68 puntos. Si comparamos este modelo con el anterior de tres variables que reduce la deviance a 48.99, encontramos que al agregar una cuarta variable, cualesquiera ella sea, no reduce significativamente más variación independiente. Esto significa que el modelo óptimo y parsimonioso para predecir eficientemente el involucramiento nodular es aquel de tres variables que incluye a X-ray + Size + log(acid).

Ciertamente es posible inspeccionar posibles interacciones entre las variables del mo-

delo que pudieran explicar otros puntos de variación de la variable dependiente, sin embargo la introducción de interacciones y otros términos no lineales complican severamente la interpretación de los coeficientes del modelo, por lo que para los efectos de este ejemplo introductorio, bastará con examinar los coeficientes del modelo aditivo que ha sido determinado como óptimo.

Tabla N° 4. Resultados de la regresión logística de involucramiento de los nódulos sobre el nivel de serum ácido fosfatasa (continua), resultados de un examen de rayos X (dummy variable) y tamaño del tumor (dummy variable).

Efecto	Coeficiente estimado	Error Standard	t de Student	p
Intercepto	-1.20	0.7162	-1.68	.094
Nivel de ácido fosfatasa	2.92	1.14	2.56	.04
Resultados de rayos X	2.06	0.796	2.58	.01
Tamaño del tumor	1.76	0.75	2.35	.02

Este modelo indica que en la escala logit, hay una relación lineal entre la probabilidad de un involucramiento nodular y el nivel de ácido fosfatasa. A mayor nivel de ácido, mayor es el nivel del involucramiento nodular, manteniendo constantes el tamaño del tumor y los resultados del examen de rayos X. En relación a los resultados negativos de los exámenes de rayos X –categoría usada como referencia comparativa en el dummy– cuando este examen resulta positivo existe un aumento de 2.06 en el logit del compromiso nodular, manteniendo constantes el tamaño del tumor y el nivel del ácido fosfatasa. El tamaño grande del tumor aumenta en 1.76 la probabilidad de involucramiento nodular, en relación a los tumores de tamaño pequeño, controlando por el nivel del ácido y los resultados del examen de rayos X.

CONCLUSIONES

El principal propósito de este artículo ha sido demostrar que el desarrollo de la computación en los últimos veinte años ha tenido un enorme impacto en la metodología de la investigación científica. Aún cuando, en general, este desarrollo tecnológico ha impulsado también avances en el campo de las técnicas de recolección de datos, el efecto más destacado ha sido ejercido sobre el análisis de datos de la investigación. La verdad es que el desarrollo de la informática ha provocado una verdadera revolución en el análisis de datos. Si se pudiera sintetizar en breves palabras este impacto, se debería decir que ahora es posible procesar, en segundos, gigantescas cantidades de información, que hace diez años implicaban meses de trabajo y enormes costos de procesamiento. La capacidad de almacenamiento de las memorias de los computadores se ha multiplicado por miles de bytes, en poco tiempo, mientras el espacio físico utilizado para almacenar esos datos es cada día más reducido. Hoy es posible procesar conjunta y simultáneamente enormes arrays de datos y variables que diez años atrás simplemente debían ser procesados en forma separada y sin ninguna posibilidad física de combinación o mezcla.

Temas de interés científico como la medición de la temporalidad o duración de los fenómenos, antes prácticamente imposibles de abordar dada la carencia de procedimientos computacionales que permitiesen realizar las iteraciones necesarias para la estimación estadística, ahora son fácilmente abordables por, al menos, una docena de paquetes estadísticos independientes. Aspectos tan sensitivos a una correcta modelización de los fenómenos como la multidimensionalidad asociada a la causalidad, tan complicados de abordar hace algún tiempo, al extremo de marcar la diferencia entre la investigación de pregrado y aquella más exigente desarrollada al nivel de postgrado, con el desarrollo contemporáneo de la computación ha llegado a ser hoy la norma general de toda investigación moderadamente seria.

En este contexto, la regresión logística fue analizada e ilustrada con el propósito de presentar una de las técnicas básicas del moder-

no análisis de datos categóricos para enfrentar los problemas del análisis multivariado de variables dependientes binarias. Haciendo uso de un paquete estadístico de considerable difusión internacional como SAS, se ha demostrado cómo se realiza la modelización estadística de variables binarias asociadas a los síntomas de la senilidad y al involucramiento nodular en pacientes con cancer prostático. El contraste de modelos jerárquicos, mediante el uso de la razón de verosimilitud, para seleccionar aquellos que incluyen términos eficientes y óptimos en reducir significativamente la variación de la variable respuesta, ha sido desarrollado apropiadamente para ilustrar los procedimientos de selección de modelos parsimoniosos. El análisis e interpretación de los coeficientes del modelo de regresión logìstica ha sido extensamente ilustrado tanto en su dimensión aditiva asociada a los logit o log-odds como en su vertiente multiplicativa asociada a la razón de odds.

Probablemente la regresión logística será, entre los modelos de análisis de datos categóricos, el equivalente a la regresión múltiple entre los modelos de regresión de variables continuas. Por tanto, es sumamente importante que todos quienes trabajan en el área de las ciencias sociales y de la salud sean capaces de incorporarla rápidamente entre sus instrumentos habituales de trabajo. De ese modo estarán avanzando efectivamente en el aspecto metodológico al adoptar una estrategia técnica de considerable flexibilidad, importancia y confiabilidad en el examen de los aspectos multidimensionales asociados especialmente a los fenómenos sociales y de salud.

BIBLIOGRAFIA

1. ABDELRAHMAN, A. & PHILIP MORGAN. Socioeconomic and institutional correlates of family formation: Khartoum, Sudan, 1945-75. Journal of Marriage and the Family 49: 401-412.
2. AGRESTI, ALAN. Categorical Data Analysis. 1st ed. New York. John Wiley & Sons, 1990, 558 pp.
3. AGRESTI, ALAN. An Introduction to Categorical Data Analysis. 1st ed. New York. John Wiley & Sons, 1996, 290 pp.
4. BEAN, FRANK *et. al.* Sociodemographic and marital heterogamy influences on the decision for voluntary sterilization. Journal of Marriage and the Family. 1987, 49:465-476.
5. BROWN, B. W. Prediction analyses for binary data. Biostatistics Casebook. New York . Wiley, 1980, 256 pp.
6. COX, D.R. The analysis of binary data. 1st ed. London. Methuen. 1970, 235 pp.
7. COLLET, D. Modelling Binary Data. 1st ed. London. Chapman & Hall, 1991, 369 pp.
8. DEMARIS, ALFRED. Logit Modelling. Practical Applications. Sage University Papers series on Quantitative Applications in the Social Sciences, 07-086, Newbury Park, CA SAGE, 1992, 85pp.
9. DOBSON, ANNETTE J. An Introduction to Generalized Linear Models. 1st ed. London, Chapman and Hall, 1990, 174 pp.
10. FIENBERG, STEPHEN. The analysis of cross-classified categorical data. 2nd ed. Cambridge, MA., MIT Press, 1980, 259pp.
11. HALLI, SHIVA S. Advanced Techniques of Population Analysis. 1st ed. New York, Plenum Press, 1992, 226 pp.
12. HIRSCH, ROBERT P. & RICHARD RIEGELMAN. Statistical First Aid. Interpretation of Health Research Data. 1st ed. Boston, Blackwell Scientific Publications, 1992, 400 pp.
13. HOSMER, DAVID W. & STANLEY LEMESHOW. Applied Logistic Regression. 1st ed. New York . John Wiley & Sons, 1989, 307 pp.
14. KALMUSS, DEBRA & JUDITH SELTZER. Continuity of marital behavior in remarriage: The case of spouse abuse. JOURNAL OF MARRIAGE AND THE FAMILY. 1986, 48:113-120.
15. LIAO, TIM F. 1 st ed. Interpreting Probability Models. Logit, Probit, and Other Generalized Linear Models. Sage University Papers series on Quantitative Applications in the Social Sciences, 07-101, Newbury Park, CA SAGE, 1994, 85 pp.
16. LINDSEY, JAMES K. Introductory Statistics. A Modelling Approach. 1st edd. New York. Oxford University Press, 1995, 214 pp.
17. NEWCOMER, SUSAN & RICHARD UDRY. Parental marital status effects on adolescent sexual behavior. Journal of Mariage and the Family. 1987. 49:235-240.
18. MACDERMOTT, NANCY & CYNTHIA WHITE. A Comparative Evaluation of Selected Statistical Software for Computing a Logistic Regression. Working Paper 94-27, Center for Demography and Ecology, University of Wisconsin-Madison, 1994, 24 pp.
19. MCLANAHAN, SARA. Family Structure and the Reproduction of Poverty. American Journal of Sociology. 1985, 90(4):873-901.

20. MERINO, JOSÉ MANUEL. Contextual effects on current use of modern contraceptive methods: Service availability of family planning and contraceptive prevalence in rural Colombia. Doctoral diss. Department of Sociology, The University of Texas at Austin. 1993, 232 pp.
21. MORGAN, PHILIP S. & JAY D. Teachman Logistic Regression: Description, Examples, and Comparisons. Journal of Marriage and the Family 50 (November 1988): 929-936.
22. RAO, K.V. Childlessness in Ontario and Quebec: Results from 1971 and 1981 census data. Can. Stud. Popul. 1987, 14:27-46.
23. RETHERFOR, D.; ROBERT, D. & MINJA, KIM. Statistical Models for Causal Analysis. 1st edd. New York. John Wiley & Sons, 1993, 258 pp.
24. ROBBINS, CYNTHIA *et. al.* Antecedents of pregnancy among unmarried adolescent. Journal of Marriage and the Family. 1985, 47:567-583.
25. RONCEK, DENNIS W. Using Logit Coefficients to Obtain the Effects of Independent Variables on Changes in Probabilities. Social Forces. December 1991, 70(2): 509-518.
26. SCHLOTZHAUER, DAVID C. Some Issues in Using Proc Logistic for Binary Logistic Regression. Technical Document. SAS Institute, Inc. 1995, 12 pp.
27. SMITH, KEN & CATHLEEN ZICK. The incidence of poverty among the recently widowed: Mediating factors in the life course. Journal of Marriage and the Family. 1986. 48:619-630.
28. STUDER, M. & A. THORTON. Adolescent religiosity and contraceptive usage. Journal of Marriage and the Family. 1987, 49(1):117-128.
29. TRUSSEL, J. & K. V. RAO. Premarital cohabitation and marital stability: A reassessment of the Canadian experience. Journal of Marriage and the Family. 1989, 51:535-544.
30. WHITE, J. Premarital cohabitation and marital stability in Canada. Journal of Marriage and the Family. 1987, 49: 641-647.
31. YAMAGUCHI, KAZUO & DENISE KANDEL. Drug use and other determinants of premarital pregnancy and its outcome: A dynamic analysis of competing life events. Journal of Marriage and the Family. 1987, 49:257-270.

APENDICE 1

Dataset sobre cinco predictores de involucramiento nodular en cáncer prostático.

Código: Id. Patient: 0-1; age: 3-4; acid level: 6-9; Xray: 11; Tumour Size: 13; Tumour grade: 15; nodal involvement: 17;

01	66	0.48	0	0	0	0	19	52	0.83	0	0	0	0	37	59	0.63	1	1	1	0
02	68	0.56	0	0	0	0	20	56	0.98	0	0	0	0	38	61	1.02	0	1	0	0
03	66	0.50	0	0	0	0	21	67	0.52	0	0	0	0	39	53	0.76	0	1	0	0
04	56	0.52	0	0	0	0	22	63	0.75	0	0	0	0	40	67	0.95	0	1	0	0
05	58	0.50	0	0	0	0	23	59	0.99	0	0	1	1	41	53	0.66	0	1	1	0
06	60	0.49	0	0	0	0	24	64	1.87	0	0	0	0	42	65	0.84	1	1	1	1
07	65	0.46	1	0	0	0	25	61	1.36	1	0	0	1	43	50	0.81	1	1	1	1
08	60	0.62	1	0	0	0	26	56	0.82	0	0	0	1	44	60	0.76	1	1	1	1
09	50	0.56	0	0	1	1	27	64	0.40	0	1	1	0	45	45	0.70	0	1	1	1
10	49	0.55	1	0	0	0	28	61	0.50	0	1	0	0	46	56	0.78	1	1	1	1
11	61	0.62	0	0	0	0	29	64	0.50	0	1	1	0	47	46	0.70	0	1	0	1
12	58	0.71	0	0	0	0	30	63	0.40	0	1	0	0	48	67	0.67	0	1	0	1
13	51	0.65	0	0	0	0	31	52	0.55	0	1	1	0	49	63	0.82	0	1	0	1
14	67	0.67	1	0	1	1	32	66	0.59	0	1	1	0	50	57	0.67	0	1	1	1
15	67	0.47	0	0	1	0	33	58	0.48	1	1	0	1	51	51	0.72	1	1	0	1
16	51	0.49	0	0	0	0	34	57	0.51	1	1	1	1	52	64	0.89	1	1	0	1
17	56	0.50	0	0	1	0	35	65	0.49	0	1	0	1	53	68	1.26	1	1	1	1
18	60	0.78	0	0	0	0	36	65	0.48	0	1	1	0							