UN ACERCAMIENTO DIDÁCTICO A LA REGRESIÓN LOGÍSTICA COMO HERRAMIENTA ESTADÍSTICA EN LAS CIENCIAS BIOLÓGICAS Y DE LA SALUD

Carlos E. Gómez
Universidad Pedagógica Nacional
Bogotá, Colombia

En este trabajo, se propone una unidad didáctica de introducción a la Regresión Logística, como aplicación a la Biología y Ciencias de la Salud y se presentan las nociones básicas de la regresión logística, acompañadas de ejemplos claros y precisos que permiten la mejor comprensión de los conceptos, y se proponen tópicos que podrían estudiarse en el curso de extensión.

INTRODUCCIÓN

La estadística cada día toma mayor importancia como aplicación en la mayoría de las ciencias naturales y de la salud. Es por esto que los estudiantes de estas carreras necesitan orientación en el manejo y análisis de datos relacionados a su área de interés. La Estadística desarrolla los conceptos básicos útiles para un buen empleo de los datos mediante el estudio de medidas estadísticas, análisis de gráficos, conceptos de probabilidad, estimación de parámetros, pruebas de hipótesis, pruebas no paramètricas, bondad de ajuste, pruebas de independencia, etc, útiles para la toma de decisiones, todo esto aplicado al diseño de experimentos, la regresión lineal, el control estadístico de calidad y otros.

Es por esta razón que en los espacios académicos programados para estadística en las diferentes carreras es indispensable tener en cuenta los anteriores temas y generar en el estudiante una motivación más para la investigación estadística, que le permita aplicar todas estas herramientas a datos de su propio conocimiento profesional. Además, debido al desarrollo de las nuevas tecnologías, es indispensable introducir en estos cursos, el aprendizaje de programas de computación especializados, como también el uso adecuado de calculadoras científicas, que permitan minimizar los cálculos de los datos involucrados y aumentar la capacidad de análisis e interpretación de los mismos, teniendo como soporte los ejes temáticos teóricos.

En este trabajo, se propone una unidad didáctica de introducción a la Regresión Logística, como aplicación a la Biología y Ciencias de la Salud y se presentan las nociones básicas de la regresión logística, acompañadas de ejemplos claros y precisos que permiten la mejor comprensión de los conceptos, y se proponen tópicos que podrían estudiarse en el curso de extensión.

GENERALIDADES

La regresión logística (**RL**) es una de las herramientas estadísticas más precisas y versátiles con las que se dispone para el análisis de datos. Su origen se remonta a la década de los sesenta, con el eminente trabajo en Epidemiología de Cornfield, Gordon y Smith (1961) acerca del riesgo de padecer una enfermedad coronaria y, ya en la forma como se conoce actualmente, con la contribución de Walker y Duncan (1967), en que se aborda el tema de estimar la probabilidad de

ocurrencia de cierto acontecimiento en función de varias variables. Su uso se universaliza y expande desde principios de los ochenta debido, especialmente, a las facilidades informáticas. En los últimos años se ha comprobado una presencia muy marcada de esta técnica, tanto en la literatura orientada a tratar temas metodológicos como en los artículos científicos biométricos.

Se considera un hecho o suceso que, a determinada altura de cierto proceso, puede ocurrir o no, por ejemplo: Un paciente hospitalizado muere o no antes de darle de alta, un sujeto operado se infecta o no durante cierto lapso postoperatorio, un proceso de producción agrícola está o no bajo control después de cierto tiempo, un diseño de experimentos funciona o no.

En estos contextos suele interesar la evaluación del efecto de uno o más antecedentes sobre el hecho de que el acontecimiento se produzca. Si se llama Y a la variable dependiente, que refleja la ocurrencia o no del suceso, esta variable es dicotòmica y se asume que puede tomar los dos valores siguientes:

$$Y = 1 \quad si \ el \ hecho \ ocurre$$

$$Y = 0 \quad si \ el \ hecho \ no \ ocurre$$

La situación más familiar es aquella en que se trata de evaluar el efecto de un solo factor, llamado X .

Tabla No 1: Condición de infectado para 40 pacientes hospitalizados según modelo de atención y edad.

Modelo Convencional		Modelo en Estudio	
Edad	Infección	Edad	Infección
34	NO	45	NO
21	NO	23	NO
54	SI	44	NO
67	NO	65	SI
32	SI	66	SI
56	SI	74	SI
76	SI	34	NO
44	NO	43	NO
34	NO	47	NO
21	NO	37	NO
48	NO	26	NO
39	NO	54	NO
22	NO	53	NO
45	NO	55	SI
65	SI	23	NO
67	SI	34	NO
22	NO	43	NO
32	NO	45	NO
21	SI	31	NO
76	SI	55	NO

A manera de ejemplo (Silva, 1995), se supone que se desea estudiar la infección hospitalaria posquirúrgica en pacientes operados de la cadera ($Y=1$ cuando el paciente se infecta a lo largo de la primera semana, $Y=0$ si no se infecta) y que se desea evaluar un nuevo modelo técnico-organizativo de la atención de enfermería que se presta a estos pacientes. Se define X_1 como una variable dicotòmica que vale 0 si el sujeto estuvo ingresado bajo el nuevo modelo y que vale 1 en caso de que haya estado atendido por el modelo convencional.

Se considera, además, que se quiere evaluar si la edad del paciente (X_2) se asocia al hecho de desarrollar una infección. Se establece que se han estudiado a 20 pacientes sujetos a cada uno de los dos regímenes de atención y que los resultados son los que se muestran en la Tabla No 1.

Una primera aproximación a la solución de estos dos problemas sería la siguiente:
Para el primero de ellos (si hay asociación entre el modelo de atención enfermera y el desarrollo de una infección) se puede resumir la información en una tabla de contingencia de dos filas y dos columnas. El resultado, en este caso, es el que se recoge en la tabla No 2

Tabla No 2 : Distribución de pacientes según modelo de atención enfermera y condición respecto de la infección.

	Infectados	No infectados	Total
Modelo Convencional	8	12	20
Modelo en Estudio	4	16	20

Se observa que la tasa de infección entre los acogidos al modelo nuevo (4/20 = 0.2), es la mitad que la correspondiente al modelo convencional (8/20 = 0.4). Sin embargo, la prueba corriente Ji–cuadrado arroja un valor observado $\chi^2_{obs}=1.90$, mucho menor que el 3.84 requerido para declarar que las tasas difieren significativamente (usando un error tipo I menor que 0.05).

En cuanto al segundo problema, una solución inmediata sería la comparación de la media de edad de los que se infectaron con la de los que no se infectaron. Estas resultaron ser 58.9 y 38.1 años respectivamente. Tratándose de tan pocas observaciones (12 infectados y 28 sanos) , probablemente la prueba estadística *t de Student* sea la más adecuada, pero no la más aceptada en estos casos, sino que se utiliza el *test no paramètrico de Kruskal-Wallis para dos grupos.* El resultado de este último arroja una clara diferencia significativa entre ambas series de edades: $\chi^2_{obs}=11.2$, mayor que 5.99 , el percentil 95 de la distribución Ji-cuadrado con dos grados de libertad que corresponde a la prueba mencionada.

Ninguna de las dos soluciones pasa por el uso de la regresión. Pero, puesto que la intención subyacente es evaluar si *Y* se modifica en dependencia de los valores asumidos por la variable independiente que se esté considerando, la idea de poner la variable *Y* en función de X_1 (o de

X_2, según el caso), puede también considerarse. Incluso, puede valorarse la posibilidad de que *Y* se ponga en función de ambas simultáneamente.
Usualmente, cuando se quiere poner una variable en función de otra (o de otras) , se acude al bien conocido recurso de la regresión lineal (simple o múltiple).
Sin embargo, si se ajusta alguna de las funciones:

$$Y = \alpha + \beta X_1 \qquad ; \quad Y = \alpha + \beta X_2$$

o, en caso en que se incluyan dos variables independientes, la función:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 ,$$

la situación que se genera es incongruente, por que?

Aunque son varias las razones que sugieren no hacer uso de este recurso en las circunstancias mencionadas, cabe mencionar la más obvia y de naturaleza menos técnica. El método usual de mínimos cuadrados, funcionará fluidamente desde el punto de vista aritmético. Pero cuando la función se evalúe para valores específicos de las variables independientes, se obtendrá un número que, salvo excepciones, será diferente de 1 y de 0 (los valores posibles de *Y*) y que, en ocasiones, estará fuera del intervalo (0,1), lo cual carece de todo sentido. Esto implica que la regresión lineal debe ser descartada como alternativa en la situación descrita. La regresión logística, en cambio, se ajusta adecuadamente a ella.

Lo que se propone mediante la regresión logística es, en principio, expresar *la probabilidad* de que ocurra el hecho en cuestión como función de ciertas variables que se presumen relevantes o influyentes. La forma analítica en que esa probabilidad se vincula con las *variables explicativas* se explica a continuación.

El caso más simple es aquel en que se incluye una sola variable independiente:

$$P(Y = 1) = \frac{1}{1 + \exp(-\alpha - \beta X)} . \qquad \textbf{(1)}$$

El caso más general es aquel en que se incluyen *k* variables:

$$P(Y = 1) = \frac{1}{1 + \exp(-\alpha - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_k X_k)} \qquad \textbf{(2)}$$

donde $\alpha, \beta_1, \beta_2, \cdots, \beta_k$ son los llamados *parámetros* del modelo, y donde *exp* denota la función exponencial. La expresión **(2)** es lo que se conoce como la *función logística* y **(1)** es su versión univariada. Al construir el modelo de regresión logística, **RL** , las variables explicativas, también

llamadas *covariables*, pueden ser de cualquier naturaleza: dicotòmicas, ordinales, continuas, o en su defecto, nominales, a estas dándoles el tratamiento de variables *Dummy*. Esta flexibilidad en cuanto a la información de entrada constituye uno de los mayores atractivos de la RL.

A continuación, se presenta una información general sobre la función logística.

FUNCIÓN LOGÍSTICA

Caso Univariado: Se considera la función $y = \frac{1}{1+\exp(-\alpha - \beta X)}$ **(3)**

Donde X es una variable cualquiera, aunque para efectos de esta aplicación se supondrá que es continua. Con frecuencia, esta función se define de la manera equivalente:

$$y = \frac{\exp(\alpha + \beta X)}{1+\exp(\alpha + \beta X)},$$

que es el resultado de multiplicar tanto el numerador como el denominador de **(3)** por $\exp(\alpha + \beta X)$.

La principal característica que vale tener en cuenta es que, para cualquier valor de X, se cumple que $0 < y < 1$, debido a que la función exponencial produce valores mayores que cero para cualquier valor. Tal circunstancia conlleva de manera natural a considerar la posibilidad de que ***y*** represente una **probabilidad.**

Otra particularidad importante es que una simple transformación de *y* produce una función lineal de X. Concretamente, aplicando las propiedades de la función exponencial y de su inversa, el logaritmo, es fácil ver que se cumple lo siguiente:

$$\ln \frac{y}{1-y} = \alpha + \beta X .$$

La transformación que atribuye a cada valor *y* el valor $\ln \frac{y}{1-y}$ es lo que se conoce como "transformación logit", y a ese número como: "el logit de *y* ", que se representa mediante *logit (y).* La función logística tiene siempre la forma de una *S* estilizada. Si $\beta > 0$, entonces la función es creciente; en caso contrario decreciente. También se puede interpretar que cuanto mayor sea $|\beta|$, mas "abrupta" es la modificación de la curva. Por ejemplo, si $\beta > 0$, a mayor valor de β, más rápidamente crece la función en la medida que aumenta X.

Por otra parte, α es tal que $\frac{1}{1+\exp(-\alpha)}$ representa la altura a la cual la curva corta al eje de ordenadas.

El gráfico siguiente muestra tres curvas correspondientes a diferentes pares de valores de α y β.

$\alpha = 2 \qquad \beta = -0.02$

$\alpha = 18 \qquad \beta = -0.08$

$\alpha = -10 \qquad \beta = 0.04$

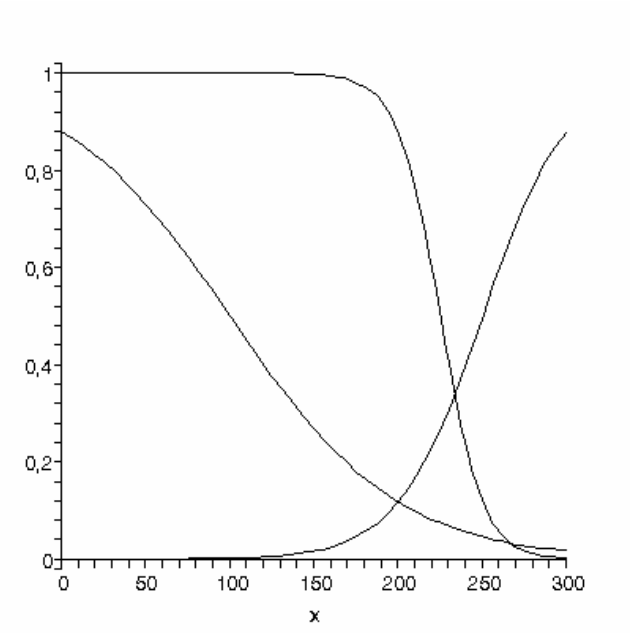


Gráfico 1: Curvas logísticas para diferentes juegos de parámetros

Obsérvese el carácter creciente de la que tiene pendiente positiva y decreciente de las otras dos. Por otra parte, analizando estas últimas, se ve que la que tiene mayor valor de α corta al eje mucho más arriba que la otra, en tanto que la que tiene mayor valor de $|\beta|$, desciende mucho más rápidamente.

El valor en que la curva cambia de concavidad (punto de inflexión, respecto del que la curva es, además, simétrica), es aquel para el cual $X = -\frac{\alpha}{\beta} : y = 0.5.$ Esto quiere decir que, si *y* representa una probabilidad, el valor de X para el cual ésta es igual a 0.5 viene dado siempre por $-\frac{\alpha}{\beta}$.

Caso Multivariado: El caso multivariado es aquel en que, en lugar de tratarse de una función de una sola variable, ***y*** se pone en función de *k* variables. Por ejemplo, para el caso *k = 3* variables, la función sería la siguiente:

$$y = \frac{1}{1 + \exp(-\alpha - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3)} \quad . \qquad \textbf{(4)}$$

Si se fijan los valores de 2 de las 3 variables explicativas, entonces **(4)** asume la forma **(3)** y queda como una función logística univariada de la variable no fijada, con la pendiente de ésta y el intercepto modificado según los valores fijados para dichas variables.

Una generalización de la función logística fue planteada por Hastie y Tibishirani (1987); consiste en la siguiente función:

$$y = \frac{1}{1 + \exp(-\alpha - f_1(X_1) - f_2(X_2) - \cdots - f_k(X_k))} \qquad \textbf{(5)}$$

donde f_i es una función cualquiera de X_i. Si la función es simplemente una homotecia de razón β_i, **(5)** se reduce a **(4)**. Dicho de otro modo, si: $f_i(X_i) = \beta_i X_i$, entonces **(5)** no es otra cosa que la función logística multivariada inicialmente definida.

Ahora, la pregunta inmediata de interés sería: ¿Es útil esta representación funcional para evaluar e interpretar problemas como los mencionados inicialmente? La respuesta no sólo es afirmativa sino que, la solución que puede darse a problemas como éstos con ayuda de la **RL** es claramente más eficiente e integral. Ejemplos y aplicaciones con datos reales es lo que se pretende que hagan los estudiantes de ciencias biológicas y de la salud

Para poder interpretar adecuadamente los coeficientes de la RL es necesario entender lo que se conoce con el término de *"odds",* que se presenta a continuación.

LOS ODDS

Los "*odds"* asociados a cierto suceso se definen como la razón entre la probabilidad de que dicho suceso ocurra y la probabilidad de que no ocurra; es decir, un número que expresa cuanto más probable es que se produzca frente a que no se produzca el hecho en cuestión.

Si se llama *E* a dicho suceso, $P(E)$ a la probabilidad de que ocurra y $O(E)$ a los *odds* que le corresponden, entonces se tiene:

$$O(E) = \frac{P(E)}{1 - P(E)}$$

A manera de ilustración, si se estima que el 75% de los pacientes que ingresan en un servicio hospitalario de quemados sobreviven, se dice que *los odds* de que un paciente genérico sobreviva son 3, (ya que 0.75 / 0.25 = 3), es decir, es 3 veces más factible que sobreviva frente a que no lo haga.

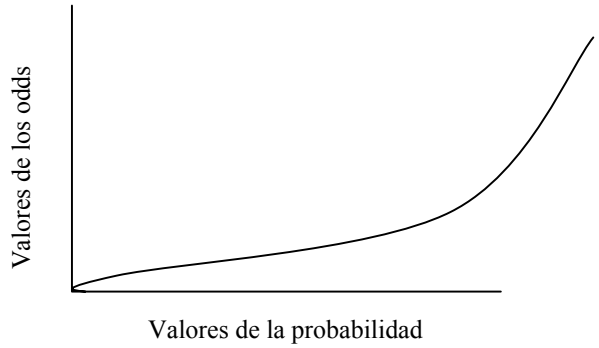
Por otra parte, conocidos *los odds*, se puede deducir la probabilidad. En general, si los *odds* de un suceso *E* es $O(E)$, entonces su probabilidad es:

$$P = \frac{O(E)}{O(E)+1},$$

por ejemplo, si se informa que *los odds* de sobrevivir que tiene un paciente operado de cáncer pulmonar son 0.4, esto equivale a decir que la probabilidad de que ese hecho ocurra es: 0.4 / 1.4 = 0.285.

De modo que ambas informaciones son equivalentes y expresan la misma noción: cuantifican cuán probable es que algo ocurra (en particular, cuál es el riesgo de un acontecimiento).Lógicamente que entre la probabilidad del suceso y los *odds* correspondientes hay una clara relación directa: si aquella aumenta, estos también lo hacen. Si $P(E) = 0$, entonces $O(E)$ también es nulo; pero en la medida que $P(E)$ tiende a la unidad, $O(E)$ tiende al infinito.

El gráfico siguiente refleja la relación existente entre ambas magnitudes:



ODDS RATIO

Se define el llamado *odds ratio* como la razón de los *odds* correspondientes a un suceso bajo cierta condición entre los que le corresponden bajo otra, esto se expresa así:

$$odds\ ratio = \frac{\dfrac{P_F(E)}{1-P_F(E)}}{\dfrac{P_{\overline{F}}(E)}{1-P_{\overline{F}}(E)}},$$

donde $P_F(E)$ y $P_{\overline{F}}(E)$ denotan respectivamente, la probabilidad de que ocurra el evento *E* cuando está presente cierta condición F y la probabilidad de que ocurra el evento *E* cuando no está presente la condición ($\overline{F}$).

Con lo anterior y observando la ecuación **(1)** definida en la sección **2**. es fácil notar que los odds del suceso *Y = 1* pueden colocarse del siguiente modo:

$$O(X)=\frac{P(Y=1)}{P(Y\neq 1)}=\frac{P(Y=1)}{1-P(Y=1)}=\exp(\alpha+\beta_1 X_1+\cdots+\beta_k X_k)\ . \qquad \textbf{(6)}$$

Ahora, se supone que se consideran dos perfiles concretos:

$$X_1^*, X_2^*, \cdots, X_k^*$$
$$X_1^\circ, X_2^\circ, \cdots, X_k^\circ$$

y que se evalúa la función **(6)** en cada uno de ellos obteniéndose $O(X^*)$ *y* $O(X^\circ)$ como los valores respectivos. Esto quiere decir que $O(X^*)$ representa los odds correspondientes al primer perfil y $O(X^\circ)$ los inherentes al segundo.
Estos odds conducen a la siguiente expresión:

$$\frac{O(X^*)}{O(X^\circ)}=\exp\left[\sum_{i=1}^{k}\beta_i\left(X_i^*-X_i^\circ\right)\right]. \qquad \textbf{(7)}$$

La fórmula **(7)** es de gran interés, pues coloca directamente una medida relativa del riesgo correspondiente a un perfil respecto de otro en términos de los parámetros de la RL.

A manera de ejemplo, se podría pensar en un proceso de producción agrícola, donde se mida la calidad de un cierto tipo de insecticida. Si se admite que la probabilidad de que el proceso se encuentre bajo control, *P(Y =1)*, esté en función de tres variables presumiblemente influyentes en este proceso, como el peso, grado de dureza y longitud; aplicando la ecuación **(7)** con *k =3* y parámetros estimados en
$\alpha=-6.614,\ \beta_1=0.075,\ \beta_2=0.312,\ \beta_3=0.018$ se responde a preguntas como la siguiente ¿Cuánto más riesgo tiene una planta con insecticidas de 10 gramos de peso, dureza de 50 y longitud de 150 mm, que uno de 20 gramos, dureza de 40 y longitud de 100mm? . Los perfiles respectivos son

$$X_1^*=10, X_2^*=50, X_3^*=150$$
$$X_1^\circ=20, X_2^\circ=40, X_3^\circ=100$$

y **(7)** queda como:

$\frac{O(X^*)}{O(X^\circ)}=\exp[0.075(10-20)+0.312(50-40)+0.018(150-100)]$, que da como resultado:

$$\frac{O(X^*)}{O(X^\circ)} = 26.31.$$

Quiere decir que la primera situación es 26.31 más riesgosa que la segunda.
Si los perfiles son iguales salvo en la i-èsima variable, se tiene:

$X_1^* = X_1^\circ, X_2^* = X_2^\circ, \cdots, X_{i-1}^* = X_{i-1}^\circ, X_{i+1}^* = X_{i+1}^\circ, \cdots, X_k^* = X_k^\circ.$
de modo que todos los sumandos de **(7)** menos el i-èsimo se anulan, y la razón de *odds* se convierte en :

$$\frac{O(X^*)}{O(X^\circ)} = \exp[\beta_i (X_i^* - X_i^\circ)]. \qquad \textbf{(8)}$$

Si finalmente, $X_i^* = X_i^\circ + 1$, entonces **(8)** se reduce a:

$$\frac{O(X^*)}{O(X^\circ)} = \exp[\beta_i]. \qquad \textbf{(9)}$$

Por otra parte, es útil afirmar que la razón de *odds* puede escribirse del modo siguiente:

$$\frac{O(X^*)}{O(X^\circ)} = \frac{P^*(Y=1)}{P^\circ(Y=1)} \frac{(1-P^\circ(Y=1))}{(1-P^*(Y=1))} \qquad \textbf{(10)}$$

donde $P^*(Y=1)$ denota $P(Y=1)$evaluado en X^* y $P^\circ(Y=1)$ denota esa misma función pero evaluada en X°.
En muchas situaciones, en especial cuando el suceso en estudio ocurre con probabilidad muy baja, el segundo factor de la derecha de la ecuación **(10)**, es decir:

$\frac{(1-P^\circ(Y=1))}{(1-P^*(Y=1))}$ es prácticamente igual a la unidad, de modo que, usando **(10)**, se puede poner:

$$\frac{O(X^*)}{O(X^\circ)} \approx \frac{P^*(Y=1)}{P^\circ(Y=1)} \qquad \textbf{(11)}$$

Por ejemplo, si $P^*(Y=1) = 0.09$ y $P^\circ(Y=1) = 0.05$, la razón de *odds* calculada mediante (10) y (11) es igual, respectivamente, a: *1.9 y 1.8*. Estos dos números son básicamente iguales, ya que, en términos prácticos, para cualquiera de ambos casos se dice que la razón de *odds* es aproximadamente igual a 2.

Para terminar, se deja abierto el problema relacionado con la construcción de la función logística a partir de observaciones reales e interpretación de las estimaciones resultantes, que implicaría profundizar en estimación de parámetros, pruebas de hipótesis, pruebas no paramètricas, bondad de ajuste, pruebas de independencia y otros temas afines para estudios caso-control.

REFERENCIAS

Anderson, T.W., "An Introduction to Multivariate Statistical Analysis", 2ed, John Wiley and Sons , New York.1984.

Dallas E. Johnson, "Métodos multivariados aplicados al análisis de datos", Internacional Thomson Editores. 2000.

Diaz, L.G., "Estadística Multivariada, Inferencia y métodos" Universidad Nacional de Colombia. 2002.

Fleis, Joseph L., "Statistical methods for rates and proportions" Ed. John Wiley. New York.1981.

Gnanadesikan, R. "Methods For Statistical Analysis of Multivariate Observations", John Wiley and Sons, New York, 1997.

Rencher, Alvin C., "Multivariate Statistical Inference and Applications", John Wiley and Sons, New York, 1998.

Silva,L.C. . ,"Excursión a la Regresión Logística en Ciencias de la Salud," Ediciones Diaz de Santos, S.A.1995.