CAPÍTULO 5

Regresión logística

5.1 Introducción

Como se expresa en la ecuación (4.1a), un modelo estadístico tiene como finalidad principal explicar el comportamiento (en términos de variabilidad) de las variables que, de acuerdo con el marco conceptual asumido por el investigador, están ligadas a un fenómeno mediante otras variables asociadas al mismo fenómeno. Un modelo está compuesto por la variable a explicar (dependiente o respuesta) y las variables explicativas (independientes o regresoras) con las cuales se pretende dar cuenta del comportamiento de la variable respuesta. El modelo se hace visible a través de una función matemática con la cual se expresan las relaciones entre las variables puestas en juego.

A continuación, se listan algunos casos que se pueden abordar con la técnica que se desarrollará de manera general en este capítulo:

- Un sujeto operado se infecta o no durante cierto lapso postoperatorio.
- Un bebé nace con malformación congénita o sin esta.
- Un paciente hospitalizado muere o no antes del alta.
- A los tres meses de vida, un niño ha dejado de lactar o aún se alimenta con leche materna.
- Un año después de una intervención quirúrgica, se ha resuelto o no el problema que la originó.
- Después de un tratamiento de quimioterapia en un paciente con cáncer de pulmón se observa alguno de los siguientes resultados sobre la enfermedad: aumento, no cambio, remisión parcial, remisión completa.

En casos como los anteriores, usualmente el interés se presta sobre la evaluación del efecto de uno o más antecedentes relacionables[1] con la ocurrencia del evento.

A diferencia del último caso, los demás eventos muestran solo dos resultados: *ocurrencia* o *no ocurrencia* de un hecho. Considere Y la variable dependiente que indica la ocurrencia o no del suceso, es decir, es una variable dicotómica. Admita que asume los valores

$$Y = 1, \text{ si el hecho ocurre,}$$
$$Y = 0, \text{ si el hecho no ocurre.}$$

El caso más sencillo trata de evaluar el efecto de un único factor X sobre una variable Y.

5.2 Modelo de regresión logística

A manera de ejemplo, considere el caso que estudia la infección hospitalaria quirúrgica en pacientes intervenidos de la cadera. De manera que $Y = 1$, cuando el paciente se infecta a lo largo de la primera semana, y $Y = 0$, si no se infecta. Se quiere evaluar un nuevo modelo técnico-organizativo del servicio de enfermería que se dispensa a estos pacientes. Sea X_1 una variable dicotómica que vale 0 si el sujeto estuvo tratado bajo el nuevo modelo y vale 1 en caso de que haya sido tratado por el modelo tradicional. Además, se quiere evaluar si la edad del paciente, X_2, se asocia al desarrollo o no de la infección.

La tabla 5.1 contiene la información sobre 40 pacientes, divididos en dos grupos de 20, cada uno de los cuales estuvo sometido a uno de los dos tipos de atención.

Tabla 5.1. Infección en pacientes hospitalizados

Atención tradicional		Atención propuesta	
Edad	Infección	Edad	Infección
34	No	45	No
21	No	23	No
54	Sí	44	No
67	No	65	Sí
32	Sí	66	Sí
56	Sí	74	Sí
76	Sí	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	Sí
65	Sí	23	No
67	Sí	34	No
22	No	43	No
32	No	45	No
21	Sí	31	No
76	Sí	55	No

Fuente: Silva (1995, p. 4).

Una primera inquietud es si existe asociación entre el modelo de atención de enfermería y el desarrollo de una infección. Este problema se resuelve como se muestra en el capítulo 2, mediante el análisis de una tabla de contingencia 2×2, mostrada en la tabla 5.2.

Tabla 5.2. Pacientes por atención e infección

Modelo de atención	Condición de infección		Total
	Infectados	No infectados	
Mod. tradicional	8	12	20
Mod. propuesto	4	16	20
Total	12	28	40

La tasa de infección entre los cobijados por el modelo de atención propuesto ($4/20 = 0.2$) es la mitad de los que corresponden al modelo tradicional ($8/20 = 0.4$). No obstante, la estadística ji-cuadrado (ecuación 2.7b) muestra un valor de $\chi_0^2 = 1.90$, el cual es menor que 3.84 (el cuantil de una ji-cuadrado con un grado de libertad y $\alpha = 0.05$); por tanto, se puede afirmar que las tasas de infección no difieren significativamente en los dos tipos de tratamientos.

Con relación a la edad, se podría comparar la media de edad de los que se infectaron con la media de la edad de los que no se infectaron. Las medias, de acuerdo con

la tabla 5.1, son 58.9 y 38.1 años, respectivamente. Tanto la prueba paramétrica *t*-Student como la no paramétrica de Wilcoxon (capítulo 11), coinciden en advertir que existe una diferencia significativa entre la edad promedio de los dos grupos.

Ninguna de las dos soluciones anteriores emplea la regresión. Es más, como la intención es evaluar si la adquisición de la infección o no, Y, es dependiente de los valores asumidos por la variable independiente tenida en cuenta, se puede considerar la variable Y en función de la variable X_1, X_2 o ambas.

La técnica estadística que se emplea para estos propósitos es la *regresión* lineal. Los modelos de regresión con los que se podría evaluar la adquisición de la infección son:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \epsilon_1, \\ Y &= \beta_0 + \beta_2 X_2 + \epsilon_2 \\ Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_3. \end{aligned} \tag{5.1}$$

Estos modelos tienen los siguientes inconvenientes (Peña, 1998, p. 501):

1. El valor esperado, evaluado para valores particulares de las variables independientes (en el i-ésimo individuo), corresponde a *la probabilidad de que la característica estudiada esté presente, esto es,* $p_i = P(Y = 1|x_{i1}, x_{i2})$. Este será un número, que salvo algunas excepciones, estará entre 0.0 y 1.0.

2. Conocidos los valores de las X, los únicos valores posibles de Y son 0.0 y 1.0; por tanto la distribución de los ϵ_i es discreta, con valores, por ejemplo, para el primer modelo de los contenidos en (5.1), $(1 - \beta_0 - \beta_1 X_1)$ y $-(\beta_0 + \beta_1 X_1)$, es decir, $(1 - p_i)$ y $(-p_i)$. Se verifica que el valor esperado de ϵ_i es

$$\mathrm{E}(\epsilon_i) = p_i(1 - p_i) + (1 - p_i)(-p_i) = 0. \tag{5.2}$$

   Por consiguiente, la variable ϵ_i tiene media cero, pero no sigue una distribución normal. La varianza de ϵ_i es

$$\mathrm{var}(\epsilon_i) = (1 - p_i)^2 p_i + (1 - p_i)p_i^2 = (1 - p_i)p_i, \tag{5.3}$$

   de manera que la varianza de los ϵ_i no es constante (hay heterocedasticidad).

En consecuencia, la regresión lineal debe ser descartada como alternativa a estas situaciones. La opción es la *regresión logística.*

Con la regresión logística se procura expresar *la probabilidad* de que ocurra el evento de interés como función de algunas variables, que desde la teoría (o la experiencia) se asumen como influyentes.

En su forma más simple el modelo logístico incluye una sola variable explicativa, por ejemplo, para X_1 es

$$\mu = E(Y) = P(Y = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1)]} \tag{5.4}$$

El siguiente es el caso más general, que involucra p-variables explicativas $X' = (1, X_1, \ldots, X_p)$,

$$\begin{aligned} P(Y = 1) &= \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)]} \\ &= \frac{1}{1 + e^{-X'\beta}} \end{aligned} \tag{5.5}$$

donde $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)'$ son los parámetros del modelo; $\exp(\cdot)$ se refiere a la función exponencial. La expresión (5.5) se conoce con el nombre de *función logística múltiple*, mientras que (5.4) corresponde a la función logística simple.

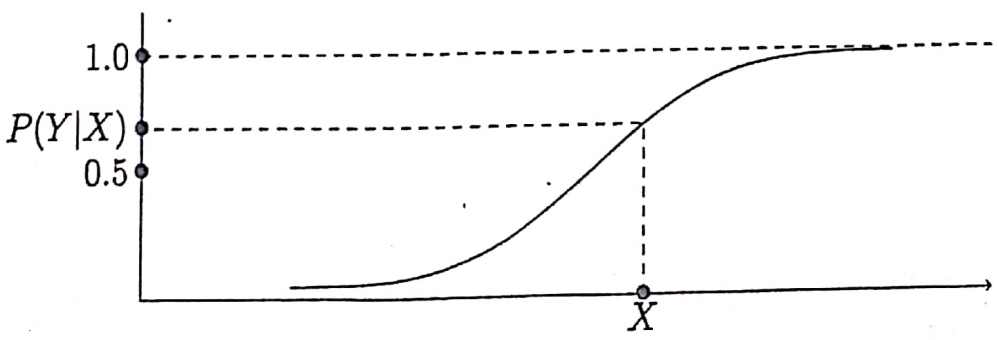
Note que

$$\begin{aligned} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p &= X'\beta \\ &= \ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) \\ &= \ln\left(\frac{\mu}{1 - \mu}\right) \\ &= g(\mu), \end{aligned}$$

es decir, la función de enlace, $g(\cdot)$, es la función logística.

Con relación a las variables explicativas del modelo de regresión logística, estas pueden ser de tipo nominal, ordinal o continuo. Este es uno de los grandes atractivos de la regresión logística. La figura 5.1 muestra la función logística univariada.

Figura 5.1. Función logística



Para ilustrar los conceptos inherentes con la regresión logística, por ahora, admita que la estimación de los parámetros del modelo

$$P(Y=1)=\frac{1}{1+\exp\left[-(\beta_0+\beta_1X_1+\beta_2X_2)\right]} \tag{5.6}$$

arroja los siguientes resultados[2]:

$$\begin{aligned} &INTERCEPTO: && \widehat{\beta}_0 = 5.3363\\ &MODELO: && \widehat{\beta}_1 = 1.4357\\ &EDAD: && \widehat{\beta}_2 = -0.1077. \end{aligned}$$

De manera que para un paciente de 54 años de edad cuya atención de enfermería sea la tradicional (es decir: $X_1 = 0$ y $X_2 = 54$), la probabilidad de contraer infecciones posteriores a la cirugía de cadera se estima remplazando las estimaciones de los β y los valores de X_1 y X_2 en (5.5), de la siguiente manera:

$$\begin{aligned} P(Y=1) &= \frac{1}{1+\exp\left[-(5.3363+1.4357(0)-0.1077(54))\right]}\\ &= \frac{1}{1+\exp\left[-(-0.4795)\right]}\\ &= 0.3824. \end{aligned}$$

Este resultado significa que aproximadamente el 38 % de los pacientes con este perfil presentará infecciones en el transcurso de la primera semana pos-cirugía.

Es importante anotar que la codificación sobre lo que representamos con $Y = 1$ o con $Y = 0$ es arbitraria e irrelevante. De forma que si se considera al revés de la optada anteriormente, las estimaciones de los parámetros solo difieren en el signo, es decir, tienen signo opuesto al que tenían antes. Para el ejemplo anterior, si se hace que $Z = 1$ corresponda a no contraer infección y $Z = 0$ a contraer infección, entonces el modelo estimado resulta

$$P(Z=1)=\frac{1}{1+\exp\left[-(-5.3363-1.4357X_1+0.1077X_2)\right]}.$$

Para calcular la probabilidad de que un individuo se infecte durante la primera semana después de la cirugía afirma que

$$\begin{aligned} P(Z=1) &= \frac{1}{1+\exp\left[-(-5.3363-1.4357(0)+0.1077(54))\right]}\\ &= \frac{1}{1+\exp\left[-(0.4795)\right]}\\ &= 0.6176, \end{aligned}$$

[2] Se obtuvieron mediante el procedimiento LOGISTIC del SAS.

que corresponde al complemento de la probabilidad de que $Y = 1$, es decir, que $P(Y = 1) = 1 - P(Z = 1) = P(Z = 0)$.

De este resultado se debe tener presente la manera en que se ha definido la variable de respuesta, pues un coeficiente con el signo positivo indica que $P(Y = 1)$ crece cuando lo hace la variable asociada al respectivo coeficiente, pero el sentido cualitativo de este hecho depende de lo que representen tanto la variable en cuestión como el evento $Y = 1$. En la siguiente sección se ampliará la interpretación de estos coeficientes.

5.3 Interpretación de los coeficientes de regresión

Una interpretación adecuada de los coeficientes β en un modelo de regresión logística va de la mano con los conceptos de: riesgo relativo, *odds* y razón *odds*, explicados en las secciones 2.5.5 y 2.5.6. En tal sentido, si se considera que las p-variables conforman un vector X, es decir, que $X = (X_1, X_2, \ldots, X_p)$, se puede probar, por las propiedades de la función exponencial, que los *odds* del evento $Y = 1$ se pueden escribir como

$$\begin{aligned} O(X) &= \frac{P(Y=1)}{P(Y \neq 1)} = \frac{P(Y=1)}{1 - P(Y=1)} \\ &= \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p). \end{aligned} \tag{5.7}$$

Suponga que se tienen dos perfiles específicos, es decir, dos individuos k y l determinados por los valores que asuman en cada una de las p-variables; estos son:

$$\begin{aligned} &\text{individuo } k: \quad X_{k1}, X_{k2}, \ldots, X_{kp} \\ &\text{individuo } l: \quad X_{l1}, X_{l2}, \ldots, X_{lp}, \end{aligned}$$

el valor de los *odds* en cada uno de ellos, de acuerdo con (5.7), son $O(X^k)$ y $O(X^l)$, respectivamente. Así, $O(X^k)$ representa los *odds* correspondientes al primer perfil y $O(X^l)$ los relacionados con el segundo. Mediante manipulación algebraica sencilla se obtiene la siguiente expresión:

$$RR = \frac{O(X^k)}{O(X^l)} = \exp\left[\sum_{i=1}^{p} \beta_i (X_{ki} - X_{li})\right], \tag{5.8}$$

donde X^k y X^l denotan el vector de observaciones para los individuos k y l, respectivamente.

La expresión (5.8) corresponde a una medida relativa del riesgo relacionada con un perfil respecto de otro en términos de los parámetros de la regresión logística.

Para el ejemplo hasta aquí tratado, la fórmula (5.8) permite responder preguntas como: ¿cuánto más riesgo tiene un sujeto de 50 años asistido por la metodología propuesta, que uno de 55 años asistido por la metodología tradicional? En este caso los perfiles son:

$$\begin{aligned} \textit{individuo}\ 1: &\quad X_{11} = 1 \text{ y } X_{12} = 50 \\ \textit{individuo}\ 2: &\quad X_{21} = 0 \text{ y } X_{22} = 55. \end{aligned}$$

Al remplazar en (5.7) el valor de las variables y las estimaciones de los parámetros, se obtiene:

$$\begin{aligned} \frac{O(X^1)}{O(X^2)} &= \exp[\widehat{\beta}_1(X_{11} - X_{21}) + \widehat{\beta}_2(X_{12} - X_{22})] \\ &= \exp[\widehat{\beta}_1(1-0) + \widehat{\beta}_2(50-55)] \\ &= 7.2001. \end{aligned}$$

Esto quiere decir que la primera situación (descrita por el perfil del individuo 1) es 7 veces más "peligrosa" que la segunda.

Si los perfiles son iguales, salvo en la i-ésima variable, en (5.7) todos los sumandos diferentes al i-ésimo se anulan, de donde resulta:

$$\frac{O(X^k)}{O(X^l)} = \exp[\beta_i(X_{ki} - X_{li})]. \tag{5.9}$$

Para el ejemplo, este caso es trivial, puesto que tan solo se tienen dos variables. Suponga que los individuos están expuestos al mismo tratamiento y que las edades son 45 y 60 años.

$$\frac{O(X^k)}{O(X^l)} = \exp[-0.1077(45-60)] = 5.0304.$$

Esto quiere decir que la primera situación es 5 veces más peligrosa (en términos de contraer infecciones) que la segunda. Aunque parezca una situación extraña, puesto que se esperaría que la metodología de tratamiento propuesta produjera mejores resultados (menos riesgo) que la tradicional. Note que tal metodología parece "favorecer" a los pacientes con edad avanzada.

Finalmente, si $X_{ki} = X_{li} + 1$ y todos los demás valores de las otras variables son iguales, (5.8) se reduce a:

$$\frac{O(X^k)}{O(X^l)} = \exp(\beta_i).$$

Por ejemplo, si los sujetos difieren tan solo en que uno se expuso al tratamiento propuesto y el otro no, entonces:

$$\frac{O(X^1)}{O(X^2)} = \exp(1.4357)$$
$$= 4.2026,$$

de donde se concluye que la metodología de asistencia propuesta aumenta el riesgo de infecciones en 4 veces, suponiendo que se han controlado las demás variables.

5.4 Construcción e interpretación de la función logística

Igual que en regresión múltiple, se deben *estimar* los parámetros β a partir de la información registrada sobre n individuos (u objetos). Tal información se dispone en notación matricial así:

$$\begin{pmatrix} Y_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Y_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}.$$

Cada fila representa el resultado de medir las variables $Y, X_1, \ldots, X_p$ en un individuo; la primera columna está compuesta por unos y ceros. La matriz tiene tantas filas como sujetos haya en la muestra. Por ejemplo, X_{35} representa la medición de la quinta variable explicativa sobre el tercer individuo de la muestra.

En regresión lineal, el método usado con más frecuencia es el de *mínimos cuadrados*, con el cual se buscan los valores de $\beta_0, \beta_1, \ldots, \beta_p$ que minimicen la suma de cuadrados de las desviaciones entre los valores observados de Y y los valores pronosticados por el modelo, es decir, se intenta encontrar los valores de los parámetros que minimicen el *error de predicción*. Bajo algunos supuestos como varianza constante, normalidad de los errores, independencia entre las variables explicativas, el método de los mínimos cuadrados produce estimadores con propiedades estadísticas deseables. Pero como se anota en las igualdades (5.2) y (5.3), las variables dicotómicas no reúnen estas propiedades; en consecuencia, se debe optar por otro procedimiento de estimación.

Dado que la variable Y sigue una distribución tipo Bernoulli (sección 1.4.2), el procedimiento adecuado de estimación es de *máxima verosimilitud* (sección 1.5.1). La solución a las ecuaciones implicadas en este proceso de optimización suministran los estimadores de los parámetros del modelo, tal explicación se sale de los propósitos de este texto; los interesados pueden consultar a Hosmer y Lemeshov (1989), Agresti