Departamento de Estadística e I.O.

Máster en Estadística Aplicada

UNIVERSIDAD
DE GRANADA

MODELOS DE RESPUESTA DISCRETA APLICACIONES BIOSANITARIAS

Tema 1
Introducción a los modelos de respuesta binaria

Profesores

Ana María Aguilera del Pino

Manuel Escabias Machuca

Título original: Modelos de Respuesta Discreta. Aplicaciones Biosanitarias. Tema 1: Introducción a los modelos de respuesta binaria

Índice general

Capítulo 1

Introducción a los modelos de respuesta binaria

Los modelos de regresión tienen como objetivo describir el efecto de una o más variables explicativas (independientes) sobre una o más variables respuesta (dependientes). En muchas aplicaciones la variable respuesta es discreta (toma pocos valores), tratándose usualmente de una variable categórica con dos o más posibles clasificaciones o niveles de respuesta. Los modelos de regresión más utilizados, en la mayoría de los campos de aplicación, para analizar este tipo de respuestas son los modelos de regresión logística (*logit*), para los que las variables explicativas pueden ser tanto cuantitativas como cualitativas.

Las pretensiones de la modelización *logit* son idénticas a las de cualquier otra técnica de regresión estadística. Se trata de encontrar el modelo más parsimonioso que se ajuste bien a los datos observados, tenga una interpretación sencilla en términos de asociación e interacción y proporcione buenas estimaciones de las probabilidades de respuesta. La diferencia fundamental entre los modelos de regresión lineal y los *logit* es que en los primeros la variable respuesta es cuantitativa y en los segundos es una variable categórica binaria o politómica.

Aunque la regresión logística es la técnica más usual para el análisis de datos de respuesta binaria, existen otros modelos alternativos, pertenecientes todos ellos a la familia de los *modelos lineales generalizados* que contiene también a otros modelos estándar de regresión como, por ejemplo, la regresión lineal y el análisis de varianza para variables respuesta continuas.

1.1. Planteamiento del problema

El objetivo es construir un modelo estadístico para estimar una variable respuesta discreta (binaria o politómica) en función de una o varias variables explicativas que podrían ser cuantitativas o cualitativas.

Comenzaremos por el caso más simple en el que se quiere explicar una variable aleatoria de respuesta binaria Y, con dos posibles categorías de respuesta (Y_1, Y_2) en función de una variable no aleatoria cuantitativa X. Ejemplos usuales de variables de respuesta binarias son los siguientes: tener una enfermedad (si, no); intención de voto (centro, no centro); opinión (a favor, en contra); etc.

Si representamos a las dos categorías de Y por los valores 0 y 1, Y tiene distribución de Bernouilli de esperanza

$$E[Y] = P[Y = 1] = p \quad (0 < p < 1).$$

Entonces, la distribución de Y en cada valor observado de X es también Bernouilli de esperanza

$$E[Y|X = x] = P[Y = 1|X = x] = p(x),$$

y varianza

$$Var[Y|X = x] = E[Y^2|X = x] - (E[Y/X = x])^2 = p(x)[1 - p(x)].$$

De este modo, $p(x)$ representa la dependencia de la probabilidad de respuesta 1 respecto de los valores de la variable explicativa.

Si denotamos por $Y(x)$ a la distribución de Y condicionada a $X = x$ $(Y|X = x)$, el paso siguiente es construir un modelo adecuado para $Y(x)$ de la forma

$$Y(x) = \text{función(parámetros, x, error)}.$$

1.2. Inviabilidad del modelo de probabilidad lineal

El modelo más sencillo para la v.a. Y en términos de X es el modelo de regresión lineal

$$Y(x) = \alpha + \beta x + \epsilon(x),$$

donde los errores $\epsilon(x)$ son variables aleatorias no observables, independientes, con esperanza cero, cuya distribución es también de Bernouilli con valores

$(1-(\alpha+\beta x))$ si $Y(x)=1$, y $-(\alpha+\beta x)$ si $Y(x)=0$, a los que corresponden las mismas probabilidades $p(x)$ y $(1-p(x))$ que la v.a. $Y(x)$.

Dado que $\epsilon(x)$ tiene esperanza cero, se tiene

$$E[\epsilon(x)] = p(x) - (\alpha+\beta x) = 0.$$

Por lo tanto, el modelo de regresión lineal es de la forma

$$E[Y|X=x] = p(x) = \alpha+\beta x. \tag{1.1}$$

y recibe el nombre de *modelo de probabilidad lineal*.

Este modelo presenta importantes defectos estructurales que le hacen inviable para explicar el comportamiento de las probabilidades de respuesta, y se enumeran a continuación:

1. Las probabilidades son valores entre cero y uno, mientras que las funciones lineales de variables cuantitativas pueden tomar valores en toda la recta real. Por lo tanto, el modelo (1.1) puede predecir valores imposibles fuera del intervalo $(0,1)$ para valores de x suficientemente pequeños o grandes. Esto se debe a que la esperanza de una variable dicotómica no puede estar explicada linealmente por una variable cuantitativa sobre un rango de valores no acotado. Por lo tanto, el modelo (1.1) sólo podría ser válido sobre un rango finito de valores de X.

2. No se satisface la condición de homocedasticidad ya que la varianza de la variable respuesta, $Var(Y|X=x) = p(x)(1-p(x))$, no es constante sobre los valores observados de X. Como consecuencia los estimadores de mínimos cuadrados ordinarios de los parámetros del modelo lineal serían insesgados pero no eficientes (no tendrían varianza mínima dentro de la clase de los estimadores lineales insesgados). Para resolver este problema y obtener estimadores más eficientes, se podrían usar mínimos cuadrados ponderados. Cada observación se ponderaría por el inverso de la varianza condicionada tomando como valor inicial $\hat{p}(x)$ el estimador de mínimos cuadrados ordinario, y usando este procedimiento iterativamente. Ésta aproximación de mínimos cuadrados reponderados iterativamente converge a los estimadores de máxima verosimilitud (MV) pero continúan las dificultades cuando $\hat{p}(x)$ se sale del intervalo [0,1].

3. Al no tener Y distribución normal, no se pueden usar las distribuciones muestrales de los estimadores de mínimos cuadrados ordinarios para hacer inferencia sobre el modelo.

4. El modelo lineal implica variaciones iguales de la probabilidad de respuesta frente a variaciones iguales de la variable explicativa. Esto no es ni mucho menos realista porque es de esperar que los cambios en X tengan menos impacto sobre p cuando la probabilidad de respuesta esté próxima a cero o a uno que cuando esté próxima a 0,5. Como ejemplo, supongamos que en un estudio epidemiológico se quiere explicar la probabilidad de desarrollar cáncer de hígado en función de la cantidad de alcohol ingerida. Lógicamente un aumento en tres cervezas en la consumición diaria influirá menos sobre esta probabilidad para un alcohólico que para una persona que se toma una cerveza diaria.

Debido a estas dificultades nos planteamos ajustar un modelo no lineal que implique una relación entre x y $p(x)$ que sea curvilínea, monótona, y acotada entre cero y uno. Las funciones de distribución de variables continuas definidas sobre toda la recta real podrían ser transformaciones adecuadas que cumplen estos objetivos. A continuación estudiaremos que tomando la función de distribución logística se obtienen los modelos de regresión logística, con la función de distribución de una normal se tienen los modelos probit y con la función de distribución de Gumbel los modelos de valores extremos.

1.3. Modelos de respuesta binaria usuales

Teniendo en cuenta lo razonado anteriormente, buscamos un modelo de la forma

$$Y(x) = F(\alpha + \beta x) + \epsilon(x)$$

con $\epsilon(x)$ vv.aa. independientes de esperanza cero, o equivalentemente

$$p(x) = F(\alpha + \beta x) \tag{1.2}$$

donde F es una función de distribución estrictamente creciente, que a su vez puede expresarse en la forma

$$F^{-1}(p(x)) = \alpha + \beta x.$$

.

1.3.1. Modelos logit

El modelo de regresión logística simple es de la forma

$$p(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{1}{1 + \exp[-(\alpha + \beta x)]}. \tag{1.3}$$

El modelo se puede escribir equivalentemente en la forma

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = \alpha + \beta x.$$

donde la transformación $\ln[p(x)/(1-p(x))]$ recibe el nombre de *logit* y $p(x)/(1-p(x))$ representa la ventaja de respuesta 1 para el valor observado x.

Características de la curva de respuesta logística

1. La curva logística representada por la ecuación (1.3) implica una relación estrictamente monótona no necesariamente creciente entre la probabilidad de respuesta y la variable explicativa que tiene forma de S y con valores en el intervalo [0,1].

2. Si $\beta > 0$, $p(x) \uparrow 1$ cuando $x \to \infty$ y $p(x) \uparrow 0$ cuando $x \to -\infty$

   Si $\beta < 0$, $p(x) \uparrow 1$ cuando $x \to -\infty$ y $p(x) \uparrow 0$ cuando $x \to \infty$.

   Esto significa que las rectas $y = 1$ e $y = 0$ son asíntotas horizontales de la curva logística. Además, $\beta > 0$ implica que la curva es creciente y $\beta < 0$ que es decreciente.

3. La tasa de cambio (crecimiento o decrecimiento) en $p(x)$ por cada unidad de cambio en x no es constante como en el caso de la regresión lineal. Efectivamente, la tasa de cambio es la pendiente de la recta tangente a la curva logística en cada punto x

   $$p'(x) = \beta p(x)(1-p(x)).$$

   Observemos que esta función depende de x y alcanza su valor máximo $p'(x) = \beta/4$ cuando $p(x) = 1/2$ que corresponde al punto de inflexión de la curva logística $x = -\alpha/\beta$. Esto quiere decir que la tasa de crecimiento o decrecimiento aumenta al aumentar $|\beta|$ y además, tiende a ser muy pequeña para valores de $p(x)$ próximos a cero o a uno.

4. Cuando el modelo *logit* (1.3) se verifica con $\beta = 0$, la curva logística se convierte en una línea recta y la variable respuesta Y es independiente de X.

5. Para mayor intuición debemos tener en cuenta que la curva logística es la función de distribución de una v.a. con distribución de probabilidad logística. Para comprobarlo, recordemos que la función de distribución

de una v.a. logística con parámetro de localización μ y parámetro de escala $\tau > 0$ es

$$F(x) = \frac{\exp\left[(x-\mu)/\tau\right]}{1+\exp\left[(x-\mu)/\tau\right]},$$

siendo una distribución simétrica con media μ y desviación estándar $\tau\pi/\sqrt{3}$.

Por lo tanto se tiene lo siguiente:

a) Si $\beta > 0$, la curva logística (1.3) es la función de distribución de una v.a. logística de parámetros $\mu = (-\alpha/\beta)$ y $\tau = 1/\beta$.

b) Si $\beta < 0$, la curva $(1 - p(x)) = 1/(1 + \exp(\alpha + \beta x))$ es la función de distribución de una v.a. logística de parámetros $\mu = (-\alpha/\beta)$ y $\tau = -1/\beta$.

1.3.2. Modelos probit

Sea Φ la función de distribución de una normal estándar (media cero y varianza uno) dada por

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt \quad \forall z \in \mathbb{R}.$$

El modelo *probit* simple es de la forma

$$p(x) = \Phi(\alpha + \beta x), \tag{1.4}$$

y se obtiene tomando como función F, en la ecuación general (1.2) de un modelo de respuesta binaria, la función de distribución Φ.

Una forma equivalente para el modelo probit es

$$\Phi^{-1}[p(x)] = \alpha + \beta x. \tag{1.5}$$

Características de la curva de respuesta probit

1. La curva del modelo probit (1.4) para $p(x)$ conlleva una relación estrictamente monótona no necesariamente creciente entre la probabilidad de respuesta y la variable explicativa, con forma de S y valores en el intervalo [0,1].

2. Si $\beta > 0$, $p(x) \uparrow 1$ cuando $x \to \infty$ y $p(x) \uparrow 0$ cuando $x \to -\infty$

   Si $\beta < 0$, $p(x) \uparrow 1$ cuando $x \to -\infty$ y $p(x) \uparrow 0$ cuando $x \to \infty$.

   Por lo tanto, las rectas $y = 1$ e $y = 0$ son asíntotas horizontales. Además, se puede comprobar fácilmente, que $\beta > 0$ implica que la curva es creciente y $\beta < 0$ que es decreciente.

3. Igual que con la curva logística, la tasa de cambio en $p(x)$ por cada unidad de cambio en x no es constante. En este caso se tiene

   $$p'(x) = \beta\Phi'(\alpha + \beta x) = \beta f(\alpha + \beta x), .$$

   siendo f la función de densidad de una v.a. con distribución normal estándar. Observemos que la tasa de cambio alcanza su valor máximo $p'(x) = \beta/\sqrt{2\pi}$ en la media de la normal estándar $\alpha + \beta x = 0$, es decir, cuando $x = -\alpha/\beta$, y $p(x) = 1/2$.

4. Cuando el modelo *probit* se verifica con $\beta = 0$, la curva de respuesta (1.4) se convierte en una línea recta y la variable respuesta Y es independiente de X.

5. Si $\beta > 0$, la curva de respuesta (1.4) del modelo probit es la función de distribución de una v.a. con distribución normal de media $-\alpha/\beta$ y desviación estándar $1/\beta$.

   Si $\beta < 0$, la curva $(1-p(x)) = 1-\Phi(\alpha+\beta x)$ es la función de distribución de una v.a. normal de media $-\alpha/\beta$ y desviación estándar $-1/\beta$.

A continuación vamos a hacer una comparación de las curvas de respuesta para los modelos logit y probit que son muy similares.

La tasa de cambio máxima de ambas curvas de respuesta se alcanza en $x = -(\alpha/\beta)$. Para el modelo logit este valor máximo es $0{,}25\beta$ mientras que para el modelo probit es aproximadamente $0{,}4\beta$, de modo que coinciden cuando el parámetro β del modelo logit es 1,6 veces el β del modelo probit.

Por otro lado, las medias de las distribuciones de probabilidad asociadas a ambas curvas de respuesta son iguales. Para $\beta > 0$, la desviación estándar de la distribución logística asociada al modelo logit es $\pi/\sqrt{3}\beta$ mientras que la de

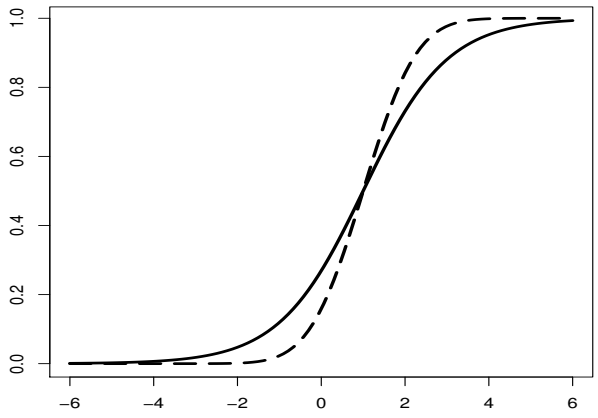

Figura 1.1: Curvas logit (linea continua) y probit (linea discontinua) con $\alpha = -1$ y $\beta = 1$

la normal asociada al modelo probit es $1/\beta$. De este modo ambas desviaciones estándar coinciden cuando el parámetro β del modelo logit es 1,8 veces el del modelo probit.

Como consecuencia, cuando tanto el modelo logit como el probit se ajustan bien, el estimador del parámetros β del modelo logit es aproximadamente 1.6-1.8 veces el del modelo probit. Finalmente, como las colas de la normal son ligeramente más estrechas que las de la distribución logística, $p(x)$ se aproxima más rápidamente a 0 y a 1 con el modelo probit que con el modelo logit.

Un caso particular de curvas de respuesta logit y probit aparecen representadas gráficamente en la Figura 1.1.

Los modelos probit se aplican con frecuencia en Toxicología para explicar la probabilidad de morir de un sujeto en términos de la dosis que se le suministra de cierta sustancia química tóxica. Sea x la dosis (o el logaritmo de la dosis) y sea la variable respuesta Y=1 si el sujeto muere. Supongamos que el sujeto tiene tolerancia T a la dosis, de modo que muere cuando la dosis suministrada es por lo menos la tolerancia ($x \geq T$). En muchos experimentos toxicológicos la distribución de la tolerancia al logaritmo de la dosis suele ser $\mathrm{N}(\mu, \sigma)$. En estos casos, el modelo para la probabilidad de morir es de la forma

$$p(x) = P(T \leq x) = \Phi[(x - \mu)/\sigma].$$

1.3.3. Modelos de valores extremos

Observemos que tanto con el modelo logit como con el modelo probit, la curva de respuesta para $p(x)$ es simétrica respecto de $p(x) = 0{,}5$. Ésto significa que el grado de aproximación de $p(x)$ a 0 y a 1 es el mismo. En este sentido, los modelos logit y probit no son adecuados para explicar probabilidades de respuesta que se alejen lentamente de 0 y se aproximen rápidamente a 1 o viceversa.

Esto justifica considerar curvas de respuesta de la forma

$$p(x) = 1 - \exp\left[-\exp\left(\alpha + \beta x\right)\right]$$

que son asimétricas respecto de p(x) = 1/2 y se alejan de 1 más bruscamente que se acercan 0.

La forma lineal equivalente a este modelo de respuesta binaria es

$$\log\left[-\log\left(1 - p(x)\right)\right] = \alpha + \beta x \tag{1.6}$$

que recibe el nombre de modelo *log-log complementario* correspondiente a la transformación del lado izquierdo de la ecuación (1.6).

El modelo alternativo en el que $p(x)$ se aleja rápidamente de 0 y se acerca lentamente a 1 es

$$p(x) = \exp\left[-\exp\left(\alpha + \beta x\right)\right], \tag{1.7}$$

o equivalentemente en forma lineal

$$\log\left[-\log\left(p(x)\right] = \alpha + \beta x,\right.$$

que recibe el nombre de modelo *log-log* de la transformación del lado izquierdo de la ecuación anterior.

Observemos que cuando el modelo log-log complementario se verifica para la probabilidad de un suceso, entonces el modelo log-log se verifica para la probabilidad de su complementario.

Características de la curva de respuesta de los modelos de valores extremos

1. Tanto para el modelo log-log complementario como para el modelo log-log, e igual que para los modelos logit y probit, las curvas de respuesta para $p(x)$ implican una relación estrictamente monótona entre la probabilidad de respuesta y la variable explicativa, con forma de S y valores en el intervalo [0,1]. De nuevo, las rectas $y = 1$ e $y = 0$ son asíntotas horizontales.

2. Para el modelo log-log complementario se tiene lo siguiente:

   Si $\beta > 0$, $p(x) \uparrow 1$ cuando $x \to \infty$ y $p(x) \uparrow 0$ cuando $x \to -\infty$. En este caso la curva es estrictamente creciente.

   Si $\beta < 0$, $p(x) \uparrow 1$ cuando $x \to -\infty$, $p(x) \uparrow 0$ cuando $x \to \infty$, y la curva de respuesta es estrictamente decreciente.

3. Para el modelo log-log se verifica

   Si $\beta > 0$, $p(x) \uparrow 0$ cuando $x \to \infty$, $p(x) \uparrow 1$ cuando $x \to -\infty$, y la curva es estrictamente decreciente.

   Si $\beta < 0$, $p(x) \uparrow 0$ cuando $x \to -\infty$, $p(x) \uparrow 1$ cuando $x \to \infty$, y la curva es estrictamente creciente..

4. La tasa de cambio en $p(x)$ para el modelo log-log complementario es

   $$p'(x) = -\beta e^{\alpha+\beta x} e^{-e^{\alpha+\beta x}} = \beta \log(1 - p(x))(1 - p(x)),$$

   que alcanza su valor máximo $p'(x) = \beta/e$ en el punto de inflexión de la curva $x = -\alpha/\beta$ al que corresponde $p(x) = 1 - e^{-1} = 0{,}6321$.

5. Análogamente la tasa de cambio en $p(x)$ del modelo log-log es

   $$p'(x) = -\beta e^{\alpha+\beta x} e^{-e^{\alpha+\beta x}} = \beta \log(p(x))p(x),$$

   que alcanza su valor máximo $p'(x) = -\beta/e$ en el punto de inflexión de la curva $x = -\alpha/\beta$ al que corresponde $p(x) = e^{-1} = 0{,}3679$.

6. De nuevo, $\beta = 0$ convierte a los modelos de valores extremos en una recta e implica que la variable respuesta Y es independiente de la variable explicativa X.

7. Para justificar la nomenclatura de modelos de valores extremos, observemos que la curva de respuesta del modelo log-log dada por (1.7) es la función de distribución de una v.a. con distribución de probabilidad de Gumbel o de valores extremos.

   Recordemos que una v.a con distribución de Gumbel de parámetros $b > 0$ y $a \in \mathbb{R}$ tiene función de distribución

   $$F(x) = \exp[-\exp[-(x - a)/b]],$$

   con esperanza $a + 0{,}577b$ y desviación estándar $\pi b/\sqrt{6}$. Por lo tanto, la curva de respuesta del modelo log-log es la función de distribución de una Gumbel de parámetros $a = -\alpha/\beta$ y $b = 1/\beta$ si $\beta > 0$ o $b = -1/\beta$ si $\beta < 0$.

Un ejemplo de curvas de respuesta de modelos de valores extremos aparece en la figura 1.2 junto a la curva logística.

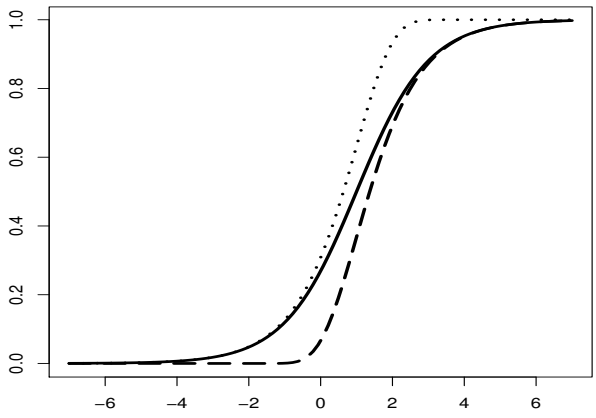

Figura 1.2: Curva de respuesta logit con $\alpha = -1$ y $\beta = 1$ (trazo continuo), curva log-log con $\alpha = 1$ y $\beta = -1$ (trazo discontinuo a rayas) y modelo log-log complementario con $\alpha = -1$ y $\beta = 1$ (trazo discontinuo punteado)

1.4. Relación con los modelos lineales generalizados

Los modelos de respuesta binaria presentados anteriormente son un caso especial de Modelos Lineales Generalizados (*GLM: Generalized Linear Models*) introducidos por Nelder y Wedderburn en 1972 y ampliamente estudiados en el libro de McCullagh y Nelder (1989).

Los modelos lineales generalizados son una amplia clase de modelos que contienen también a los modelos lineales usuales. A groso modo, un modelo lineal generalizado para una variable de respuesta aleatoria, en función de los valores observados de un conjunto de variables explicativas no aleatorias, no es otra cosa que un modelo lineal para una transformación de los valores esperados de la variable respuesta.

Si denotamos por $\mu(x_1, \ldots, x_R) = E[Y|X_1 = x_1, \ldots, X_R = x_R]$ a la esperanza de la variable respuesta Y en cada conjunto de valores observados $(x_1, \ldots, x_R)$ de las variables explicativas $(X_1, \ldots, X_R)$, un modelo lineal generalizado es de la forma

$$G[\mu(x_1, \ldots, x_R)] = \alpha + \sum_{j=1}^{R} \beta_j x_j.$$

Un modelo lineal generalizado queda así especificado mediante tres com-

ponentes

1. *Componente aleatoria:* distribución de probabilidad de la variable respuesta Y que pertenece a la familia exponencial natural. La distribución Normal y la Binomial son ejemplos de distribuciones de esta familia.

2. *Componente sistemática:* función lineal de las variables explicativas que se usa como predictor lineal.

3. *Ligadura o enlace:* función G que describe la relación funcional entre la componente sistemática y el valor esperado de la componente aleatoria.

En resumen, un GLM es un modelo lineal para una transformación de la media de una variable con distribución en la familia exponencial natural. Por lo tanto, los modelos de regresión lineal para la esperanza de una variable respuesta con distribución Normal son modelos lineales generalizados con ligadura la función identidad. Los tres tipos de modelos de respuesta binaria estudiados en este tema pueden verse también como modelos GLM cuya función ligadura es la inversa de una función de distribución F estrictamente creciente asociada a una variable aleatoria continua definida sobre la recta real. Estos modelos son de la forma

$$F^{-1}[p(x)] = \alpha + \beta x.$$

En la siguiente tabla aparece un resumen de modelos de regresión estándar que pueden verse como modelos GLM:

Modelos	C. aleatoria	Ligadura	C. sistemática
Reg. Lineal	Normal	Identidad	Continua
ANOVA	Normal	Identidad	Categórica
ANCOVA	Normal	Identidad	Mixta
Reg. Logística	Bernouilli	Logit	Mixta
Probit	Bernouilli	Inv. f.d.d. N(0,1)	Mixta
Valores Extremos	Bernouilli	Log-log	Mixta
Log-Lineales	Poisson	Logaritmo	Categórica
Respuesta Multinomial	Multinomial	Logit Generalizados	Mixta

Para la mayoría de los modelos GLM, la log-verosimilitud es estrictamente cóncava lo que implica la existencia y unicidad de los estimadores máximo verosímiles (MV) de los parámetros del modelo. Dichos estimadores MV se calculan mediante un algoritmo iterativo que usa una versión generalizada de mínimos cuadrados y se llama codificación de Fisher (*Fisher scoring*). En el caso de modelos GLM con ligadura canónica (transforma la esperanza en el

parámetro natural de la distribución exponencial) este algoritmo se simplifica en el algoritmo iterativo de Newton-Raphson. Como la transformación logit es el parámetro natural de la familia exponencial de la distribución de Bernouilli, la estimación MV de los parámetros del modelo logit se obtienen de forma sencilla mediante el algoritmo de Newton-Raphson mientras que para los modelos probit y de valores extremos la estimación MV de sus parámetros es más compleja y se lleva a cabo mediante el procedimiento de codificación de Fisher.

Otra justificación interesante para el uso de la función logística en lugar de otras funciones de distribución es la siguiente: Si X es una v.a. cuya distribución de probabilidad condicionada a Y=i (i=0,1) es $\mathrm{N}(\mu_i, \sigma^2)$ entonces, del teorema de Bayes, se deduce que

$$p(x) = P[Y = 1|X = x] = [\exp(\alpha + \beta x)]/[1 + \exp(\alpha + \beta x)],$$

con $\beta = (\mu_1 - \mu_0)/\sigma^2$.

1.5. Aplicaciones en Epidemiología

Hoy día los datos cualitativos son muy abundantes en cualquier disciplina por lo que los modelos de respuesta binaria tipo *logit*, en cuyo estudio nos centraremos, pueden ser convenientemente adaptados para explicar las probabilidades de respuesta de interés en cada campo. Sin embargo, existen campos de aplicaciones específicas del análisis de datos categóricos como la Epidemiología y la Medicina que requieren familiarizarse con una terminología propia que pasamos a introducir brevemente.

Aunque antes del siglo XX ya se realizaron algunos estudios epidemiológicos ha sido en los últimos 20 años cuando, gracias en parte a la contribución de la Estadística, ha empezado a tomar forma un cuerpo de principios sistematizado con el que analizar dichos datos.

Uno de los primeros estudios epidemiológicos serios de larga duración fue iniciado en el año 1949 por el cardiólogo Framinghan para estudiar los factores de riesgo de la enfermedad cardiovascular. Gracias a este estudio se ha podido comprender la etiología de este tremendo problema de salud pública y se han sentado las bases prácticas para su prevención.

La Epidemiología trata los patrones de distribución de las enfermedades en las poblaciones humanas, así como los factores que influyen en esos patrones. A diferencia de otras ciencias que también estudian la enfermedad, como la Medicina, la Epidemiología se centra en la ocurrencia de los procesos patológicos en lugar de en el resultado. La ocurrencia de una enfermedad se mide mediante las tasas de incidencia y de prevalencia.

La Epidemiología tiene como principal objetivo estimar el efecto que tiene la exposición a determinados factores de riesgo sobre el padecimiento de cierta enfermedad (problema de salud), controlando a su vez otras variables que puedan confundir o modificar dicho efecto.

Estudios epidemiológicos muy divulgados actualmente son aquellos que pretenden explicar la probabilidad de padecer cáncer de pulmón en función del número de cigarrillos consumidos. En este caso el hábito de fumar es un factor de riesgo considerándose como individuos expuestos los fumadores. Aparte del factor de riesgo hipotetizado cuyo efecto sobre la enfermedad se quiere estimar, existen otras variables que representan rasgos fundamentales de los individuos de la población (edad, sexo, alimentación, etc) y pueden distorsionar el efecto de la exposición de interés sobre la enfermedad. Estas otras variables deben ser controladas en cualquier estudio epidemiológico mediante un diseño estratificado y son de dos tipos: factores de confusión y factores modificadores de efecto.

Un factor de confusión es comunmente una variable que está relacionada al mismo tiempo con la enfermedad y con la exposición. Un factor modificador de efecto es una variable que puede cambiar el grado de asociación entre la enfermedad y la exposición. Esto significa que la asociación entre la enfermedad y la exposición variará en los distintos niveles de una variable modificadora de efecto mientras que será constante en los estratos asociados a un factor de confusión.

La investigación de la etiología de una enfermedad se suele llevar a cabo en tres etapas:

1. El clínico hace una observación o hipótesis respecto a la causa (factor de riesgo) de la enfermedad basándose en su experiencia.

2. Se lleva a cabo un estudio epidemiológico para cuantificar estadísticamente el efecto de la exposición sobre el padecimiento de la enfermedad.

   La calidad de la estimación dependerá fundamentalmente del diseño realizado para la recogida de una muestra no sesgada de individuos, a los que se observarán las variables de interés para después hacer inferencia estadística. A continuación se presentan los diseños muestrales utilizados en los estudios epidemiológicos que son los usuales en los estudios médicos en los que se pretende determinar el efecto de una variable explicativa (factor de riesgo o exposición) sobre una variable respuesta (padecimiento de la enfermedad).

   a) *Estudios prospectivos:* el investigador selecciona una muestra aleatoria de individuos que, después de ser observada durante cierto

tiempo, se clasifica según los niveles de la variable respuesta. En este caso las variables explicativas (factores de riesgo, de confusión y modificadores de efecto) que pueden tener relación con la respuesta son medidas prospectivamente antes de la ocurrencia de la respuesta de interés (enfermedad).

Hay dos tipos de estudios prospectivos,

1) *Estudios de cohorte* (*cohort studies*): el investigador selecciona aleatoriamente una muestra de individuos (cohorte) que hacen su propia elección sobre el grupo de la variable explicativa al que se unen, y después de un periodo de tiempo fijo se observa su respuesta.
2) *Ensayos clínicos* (*clinical trials*): el investigador selecciona aleatoriamente los individuos de cada grupo de interés definido por la variable explicativa.

Consideremos, por ejemplo, un estudio epidemiológico en el que se pretende estudiar la relación entre fumar y padecer cáncer de pulmón. Un estudio prospectivo consistiría en tomar una muestra de personas libres de cáncer y seguirla durante los próximos 30 años observando después de este tiempo las tasas de incidencia de cáncer entre fumadores y no fumadores. Si el estudio es de cohorte los individuos deciden por si mismos si fuman o no y el investigador observa simplemente quien desarrolla la enfermedad después de los 30 años, mientras que si se trata de un ensayo clínico el investigador selecciona aleatoriamente los fumadores y no fumadores.

b) *Estudios transversales* (*cross-sectional studies*): una muestra aleatoria de individuos se clasifica simultáneamente según el nivel de la variable explicativa al que pertenece y su respuesta actual. En el estudio del efecto de fumar sobre el desarrollo de cáncer de pulmón se trataría de tomar una muestra aleatoria de individuos para los que se observa su condición de fumador y si padecen actualmente la enfermedad.

c) *Estudios retrospectivos de casos y controles* (*case-control studies*): consisten en tomar una muestra de individuos en cada nivel de la variable respuesta para los que se investiga retrospectivamente en el pasado el nivel de las variables explicativas de interés.

En el estudio epidemiológico sobre el cáncer de pulmón consistiría en tomar una muestra de enfermos (casos) y otra de individuos sanos (controles) para los que se estudia si han sido o no fuma-

dores. En estos estudios se regresa al tiempo de exposición de la enfermedad que ya ha ocurrido de modo que la variable respuesta (padecimiento de la enfermedad) deja de ser aleatoria y, como se verá más adelante, no tendrá sentido estimar el riesgo relativo de padecer la enfermedad entre individuos expuestos y no expuestos al factor de riesgo.

De los estudios presentados anteriormente, los de cohorte, los transversales y los de casos y controles son observacionales mientras que los ensayos clínicos son experimentales porque es el investigador quien decide que individuos pertenecen a los grupos definidos por las variables explicativas. Los estudios retrospectivos y los transversales tienen la ventaja de ser más rápidos que los prospectivos porque los datos se consiguen en el momento. Además, los estudios de casos y controles son más útiles para detectar las causas de enfermedades poco comunes. Sin embargo el sesgo es mucho mayor poque los datos no son totalmente aleatorios y suele ser usual realizar una elección inapropiada de los controles, como el caso de individuos hospitalizados cuyas características afecten a los resultados. La principal dificultad de los estudios prospectivos es el seguimiento continuado de la cohorte.

3. Una vez identificados los factores de riesgo, se diseña un ensayo de intervención experimental para comprobar si la modificación de tales factores en los enfermos va seguida de una reducción en el padecimiento.

   El diseño muestral utilizado en esta última etapa suele ser el *ensayo clínico* cuyo propósito inicial no es identificar la etiología de la enfermedad sino determinar si un tratamiento médico es superior a otro. Para ello se suelen seleccionar dos muestras aleatorias de pacientes asignando el tratamiento a un grupo (casos) y un placebo u otro tratamiento al otro grupo (controles). Posteriormente se observa la respuesta al tratamiento.

   Para evitar que los datos recogidos sean sesgados los individuos deben ser asignados aleatoriamente a los dos grupos. Este método de asignación aleatoria ha suscitado una gran oposición entre los médicos debido al problema ético de permitir que un suceso aleatorio determine el tratamiento del paciente. Otro procedimiento habitual es comparar un nuevo tratamiento con otro ya experimentado utilizando como controles de los pacientes que reciben el nuevo tratamiento los que recibieron el tratamiento antiguo.

   La información obtenida de estos estudios suele ser muy valiosa en Epidemiología para confirmar las relaciones etiológicas sugeridas por

los estudios observacionales retrospectivos y prospectivos. Por ejemplo, el conocer que los programas para que dejen de fumar las mujeres embarazadas son efectivos en la prevención de nacimientos de niños con bajo peso, añade un fuerte apoyo para la conclusión de que el fumar durante el embarazo es una causa de bajo peso al nacer.

Independientemente del diseño realizado, el análisis estadístico de la asociación entre la exposición a un factor de riesgo y el desarrollo de una enfermedad se lleva a cabo en tres fases diferentes:

1. *Análisis simple:* se parte de una tabla de contingencia 2×2 que representa por filas al factor de riesgo (individuos expuestos y no expuestos), y por columnas a la enfermedad (enfermos, no enfermos). El test exacto de Fisher y el test chi-cuadrado serán usados para determinar si hay asociación entre el factor de riesgo y la enfermedad. El grado de asociación se estudiará mediante el cociente de ventajas (estudios de retrospectivos de casos y controles) o el riesgo relativo (estudios prospectivos y transversales). El interés principal en Epidemiología será la construcción de intervalos de confianza sobre estas medidas.

2. *Análisis estratificado:* cuando se sospecha que existen otras variables que pueden confundir o modificar el efecto entre la exposición y el desarrollo de la enfermedad, suele ser usual estratificar los datos según los niveles de dichas variables. En el caso de controlar un tercera variable se dispone de una tabla 2×2 como las del análisis simple en cada nivel de dicha variable (estrato). El primer paso será contrastar si el efecto de la exposición sobre la enfermedad es el mismo para todos los niveles del tercer factor (ausencia de interacción). Por ejemplo, determinar si el riesgo de cáncer de pulmón para los fumadores es igual en todos los grupos de edad. Si se acepta un efecto uniforme (factor de confusión) el siguiente paso es estimarlo como una medida ponderada (*pooling*) de los efectos en cada tabla parcial (estrato). Para este fín se estudirá más adelante la metodología de Mantel-Haenszel que proporciona un estimador de la razón de ventajas común. Cuando el efecto no es uniforme (interacción), la variable usada en la estratificación deja de ser un factor de confusión para convertirse en un factor modificador de efecto.

3. *Análisis multivariante:* en el caso de controlar en el estudio más de una variable, el análisis simple y el estratificado presentan importantes deficiencias como el tener que categorizar las variables continuas con la pérdida de información que esta práctica conlleva. La metodología

a seguir en este caso es la identificación y estimación de modelos estadísticos de respuesta binaria como los logit que expliquen el efecto de las variables sobre la incidencia de la enfermedad de forma precisa. En el caso de covariables que son factores de confusión su interacción con la exposición debe ser nula. Si se trata de modificadores de efecto los modelos deben incluir un término de interacción entre el factor modificador de efecto y la exposición.

1.6. Cociente de ventajas y riesgo relativo

Recordemos que la ventaja de respuesta $Y = 1$ para el valor observado $X = x$ viene dado por el cociente $p(x)/(1 - p(x))$. Entonces, se define el riesgo relativo de respuesta $Y = 1$ (padecer la enfermedad) para dos valores distintos x_1 y x_2 de la variable explicativa X, como

$$R_{12} = \frac{p(x_1)}{p(x_2)}.$$

Por otro lado, el cociente de ventajas (*odd ratio*) de respuesta $Y = 1$ dados dos valores distintos x_1 y x_2 de la variable explicativa X, es de la forma

$$\theta_{12} = \frac{\dfrac{p(x_1)}{1 - p(x_1)}}{\dfrac{p(x_2)}{1 - p(x_2)}}.$$

Observemos que en el caso de estudios retrospectivos de casos y controles el riesgo relativo de respuesta no tiene sentido ya que la variable de respuesta Y deja de ser aleatoria (el número de enfermos y no enfermos está fijado por el diseño) y es la variable explicativa X la que pasa a ser aleatoria. Sin embargo, los modelos *logit* para explicar la probabilidad de padecer la enfermedad en función del factor de riesgo seguirán teniendo sentido gracias a su interpretación en términos de cocientes de ventajas y a la siguiente relación de estos últimos con el riesgo relativo:

$$\theta_{12} = R_{12} \times \frac{1 - p(x_2)}{1 - p(x_1)},$$

de modo que cuando la probabilidad de respuesta $Y = 1$ es muy próxima a cero, el riesgo relativo puede ser aproximado mediante el cociente de ventajas que siempre se puede calcular tanto en estudios prospectivos como retrospectivos gracias a su simetría.

Consideremos, para exponer esta propiedad de simetría del cociente de ventajas, un estudio epidemiológico con un factor de riesgo binario X (0: no expuestos a la enfermedad, 1: expuestos) que incide sobre el padecimiento de determinada enfermedad Y (0: No, 1: Si). En este caso el cociente de las ventajas de padecer la enfermedad para los individuos expuestos respecto de los no expuestos es

$$\theta = \frac{\dfrac{P(Y=1/X=1)}{P(Y=0/X=1)}}{\dfrac{P(Y=1/X=0)}{P(Y=0/X=0)}} = \frac{\dfrac{P(X=1/Y=1)}{P(X=0/Y=1)}}{\dfrac{P(X=1/Y=0)}{P(X=0/Y=0)}},$$

coincide con el cociente de las ventajas de haber estado expuesto a la enfermedad para los individuos que la padecen con respecto a los que no la padecen y tiene, por lo tanto, el mismo valor tanto cuando X es la variable aleatoria (estudios retrospectivos de casos y controles) como si lo es Y (estudios prospectivos).

Para los estudios epidemiológicos de este tipo, los datos se suelen representar en forma de tablas de contingencia 2×2 como la siguiente

	$Y=1$	$Y=0$	
$X=1$	n_{11}	n_{12}	$n_{1\bullet}$
$X=0$	n_{21}	n_{22}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	n

La estimación máximo-verosímil (MV) del cociente de ventajas de una tabla 2×2 de este tipo viene dada por

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \hat{R}\frac{\dfrac{n_{22}}{n_{2.}}}{\dfrac{n_{12}}{n_{1.}}},$$

definiendo el riesgo relativo muestral como $\hat{R} = \dfrac{\dfrac{n_{11}}{n_{1.}}}{\dfrac{n_{21}}{n_{2.}}}$.

Ambas medidas de asociación muestrales tendrán valores muy parecidos cuando el porcentaje de valores con $Y = 1$ sea muy pequeño tanto para los expuestos como los no expuestos. Para un estudio más detallado sobre el cociente de ventajas y riesgo relativo asociado a una tabla de contingencia 2×2 el lector interesado puede ver Aguilera (2000).