Ajuste de curvas y el método de mínimos cuadrados

CAPÍTULO 13

RELACIÓN ENTRE VARIABLES

En la práctica es frecuente encontrar una relación entre dos (o más) variables. Por ejemplo, el peso de hombres adultos depende en cierto grado de su estatura, las circunferencias de los círculos dependen de su radio, y la presión de una masa de gas depende de su temperatura y volumen.

Entonces, es mejor expresar esta relación en forma matemática, lo cual sucede determinando una ecuación que enlaza las variables.

AJUSTE DE CURVAS

Para determinar una ecuación que relacione variables, un primer paso es recolectar datos que muestran los valores correspondientes de las variables en consideración. Por ejemplo, supóngase que X y Y denotan la estatura y el peso de hombres adultos, respectivamente; entonces, una muestra de N individuos revelaría las estaturas $X_1, X_2, \ldots, X_N$, así como los pesos correspondientes $Y_1, Y_2, \ldots, Y_N$.

El siguiente paso es graficar los puntos $(X_1,Y_1), (X_2,Y_2), \ldots, (X_N,Y_N)$ en un sistema rectangular de coordenadas. El conjunto de puntos resultante suele denominarse *diagrama de dispersión.*

A partir del diagrama de dispersión es posible visualizar una curva suave que se aproxima a los datos. Tal curva se denomina *curva de aproximación*. Por ejemplo, en la figura 13-1, los datos parecen aproximarse bien a una línea recta, por lo que se dice que hay una *relación lineal* entre las variables. Sin embargo, en la figura 13-2, aunque existe una relación entre las variables, ésta no es lineal, por lo que se le conoce como *relación no lineal.*

El problema general para encontrar ecuaciones de curvas de aproximación que se ajusten a conjuntos de datos se denomina *ajuste de curvas.*

FIGURA 13-1

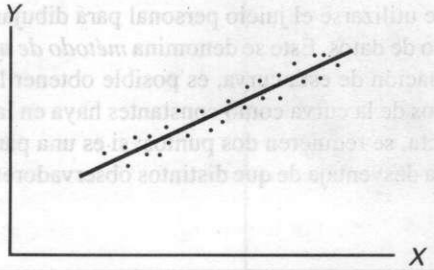
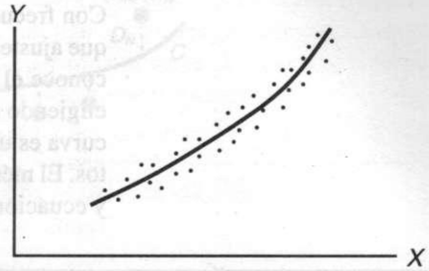


FIGURA 13-2



ECUACIONES DE CURVAS DE APROXIMACIÓN

A continuación se presenta una lista de varios tipos de curvas de aproximación y sus ecuaciones, con el propósito de tener una referencia. Todas las letras, excepto X y Y representan constantes. Las variables X y Y se conocen como *variable independiente* y *variable dependiente*, respectivamente, aunque estos papeles pueden intercambiarse.

Línea recta	$Y = a_0 + a_1X$	*(1)*
Parábola o curva cuadrática	$Y = a_0 + a_1X + a_2X^2$	*(2)*
Curva cúbica	$Y = a_0 + a_1X + a_2X^2 + a_3X^3$	*(3)*
Curva cuártica	$Y = a_0 + a_1X + a_2X^2 + a_3X^3 + a_4X^4$	*(4)*
Curva de grado n	$Y = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$	*(5)*

Las partes derechas de las ecuaciones se denominan *polinomios* de primero, segundo, tercero, cuarto y n-ésimo grados, respectivamente. Las funciones definidas por las primeras cuatro ecuaciones se llaman funciones *lineal*, *cuadrática*, *cúbica* y *cuártica*, en ese orden.

Las siguientes son algunas otras de las muchas ecuaciones usadas en la práctica con frecuencia:

Hipérbola	$Y = \dfrac{1}{a_0 + a_1X}$ o $\dfrac{1}{Y} = a_0 + a_1X$	*(6)*
Curva exponencial	$Y = ab^X$ o $\log Y = \log a + (\log b)X = a_0 + a_1X$	*(7)*
Curva geométrica	$Y = aX^b$ o $\log Y = \log a + b(\log X)$	*(8)*
Curva exponencial modificada	$Y = ab^X + g$	*(9)*
Curva geométrica modificada	$Y = aX^b + g$	*(10)*
Curva de Gompertz	$Y = pq^{b^X}$ o $\log Y = \log p + b^X(\log q) = ab^X + g$	*(11)*
Curva de Gompertz modificada	$Y = pq^{b^X} + h$	*(12)*
Curva logística	$Y = \dfrac{1}{ab^X + g}$ o $\dfrac{1}{Y} = ab^X + g$	*(13)*
	$Y = a_0 + a_1(\log X) + a_2(\log X)^2$	*(14)*

Para decidir qué curva debe utilizarse es necesario obtener diagramas de dispersión de variables transformadas. Por ejemplo, si un diagrama de dispersión de $\log Y$ contra X muestra una relación lineal, la ecuación tiene la forma *(7)*, mientras que si $\log Y$ contra $\log X$ indica una relación lineal la ecuación es de la forma *(8)*. A menudo se usa papel milimétrico para facilitar la decisión sobre cuál curva utilizar. El papel para graficar que contiene una escala dividida en forma logarítmica se conoce como *papel gráfico semilogarítmico* (o *semilog*), y aquel con las dos escalas dividas en forma logarítmica se llama *papel gráfico log-log*.

MÉTODO DE AJUSTE DE CURVAS A MANO

Con frecuencia puede utilizarse el juicio personal para dibujar una curva de aproximación que ajuste un conjunto de datos. Éste se denomina *método de ajuste de curvas a mano*. Si se conoce el tipo de ecuación de esta curva, es posible obtener las constantes de la ecuación eligiendo tantos puntos de la curva como constantes haya en la ecuación. Por ejemplo, si la curva es una línea recta, se requieren dos puntos; si es una parábola, se necesitan tres puntos. El método tiene la desventaja de que distintos observadores obtendrán diferentes curvas y ecuaciones.

LA LÍNEA RECTA

El tipo más simple de curva de aproximación es una línea recta, cuya ecuación puede expresarse

$$Y = a_0 + a_1 X \tag{15}$$

Dados cualesquiera dos puntos (X_1,Y_1) y (X_2,Y_2) en la recta, es posible determinar las constantes a_0 y a_1. La ecuación resultante de la recta se expresaría así:

$$Y - Y_1 = \left(\frac{Y_2 - Y_1}{X_2 - X_1}\right)(X - X_1) \quad \text{o} \quad Y - Y_1 = m(X - X_1) \tag{16}$$

donde $$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

se llama la *pendiente* de la recta y representa el cambio en Y, dividido entre el cambio correspondiente en X.

Cuando la ecuación se escribe en la forma (*15*), la constante a_1 es la pendiente m. La constante a_0, que es el valor de Y cuando $X = 0$, se denomina la *intersección en Y*.

EL MÉTODO DE MÍNIMOS CUADRADOS

Para evitar el juicio personal en la construcción de rectas, parábolas u otras curvas de aproximación para ajustar los conjuntos de datos, es necesario tener una definición de una "recta de mejor ajuste", "parábola de mejor ajuste", etcétera.

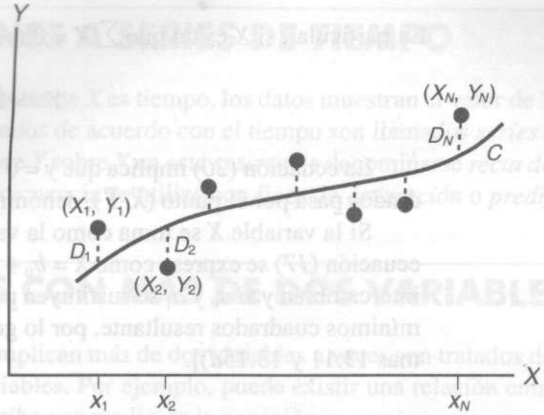
Para lograr tal definición, considérese la figura 13-3, en donde los datos están dados por los puntos (X_1,Y_1), (X_2,Y_2),..., (X_N,Y_N). Para un valor determinado de X, por ejemplo X_1, habrá una diferencia entre el valor Y_1 y el valor correspondiente deducido a partir de la curva C. Como se muestra en la figura, esta diferencia se simboliza con D_1 y se conoce como una *desviación*, un *error* o un *residual*, y puede ser positiva, negativa o cero. De manera similar, se obtienen las desviaciones D_2,..., D_N correspondientes a los valores X_2,..., X_N.

Una medida de la "bondad de ajuste" de la curva C de los datos está proporcionada por la cantidad $D_1^2 + D_2^2 + \cdots + D_N^2$. Si ésta es pequeña, el ajuste es bueno; si es grande, el ajuste es malo. Por lo tanto, se tiene la siguiente

Definición: De todas las curvas que se aproximan a un conjunto de datos definidos por puntos, la curva que tiene la propiedad de que $D_1^2 + D_2^2 + \cdots + D_N^2$ es un mínimo se denomina *curva de ajuste óptimo*.

Se dice que una curva con esta propiedad se ajusta a los datos en el *sentido de mínimos cuadrados* y se le llama *curva de mínimos cuadrados*. Entonces, una recta con esta propiedad se denomina *recta de mínimos cuadrados*, una parábola con esta propiedad se denomina *parábola de mínimos cuadrados*, etcétera.

FIGURA 13-3



Es habitual emplear la definición anterior cuando X es la variable independiente y Y es la variable dependiente. Si X es la variable dependiente, la definición se modifica pues en este caso se consideran desviaciones horizontales en lugar de desviaciones verticales, que es lo mismo que intercambiar los ejes X y Y. Estas dos definiciones generalmente conducen a curvas diferentes de mínimos cuadrados. A menos que se especifique lo contrario, se debe considerar Y como la variable dependiente y X como la variable independiente.

Es posible definir otra curva de mínimos cuadrados si se toman en cuenta distancias perpendiculares a partir de cada uno de los puntos de la curva, en lugar de distancias horizontales o verticales; sin embargo, esto no suele utilizarse.

LA RECTA DE MÍNIMOS CUADRADOS

La recta de mínimos cuadrados que se aproxima al conjunto de puntos (X_1,Y_1), (X_2,Y_2),..., (X_N,Y_N) tiene por ecuación

$$Y = a_0 + a_1 X \tag{17}$$

donde las constantes a_0 y a_1 se determinan resolviendo simultáneamente las ecuaciones

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{18}$$

denominadas *ecuaciones normales para la recta de mínimos cuadrados* (*17*). Las constantes a_0 y a_1 de las ecuaciones (*18*) pueden calcularse a partir de las fórmulas

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \tag{19}$$

Las ecuaciones normales (*18*) son fáciles de recordar si se observa que la primera ecuación puede obtenerse sumando en ambos lados de (*17*), [es decir, $\sum Y = \sum (a_0 + a_1 X) = a_0 N + a_1 \sum X$], mientras que la segunda ecuación resulta multiplicando primero ambos lados de (*17*) por X y sumando después [por ejemplo, $\sum XY = \sum X(a_0 + A_1 X) = a_0 \sum X + a_1 \sum X^2$]. Obsérvese que no es una consecuencia de las ecuaciones normales, sino sólo un medio para recordarlas. Nótese también que en las ecuaciones (*18*) y (*19*) se utilizó la notación abreviada $\sum X$, $\sum XY$, etcétera, en lugar de $\sum_{j=1}^{N} X_j$, $\sum_{j=1}^{N} X_j Y_j$, etcétera.

El trabajo requerido para encontrar una recta de mínimos cuadrados puede simplificarse algunas veces si se transforman los datos de manera que $x = X - \bar{X}$ y $y = Y - \bar{Y}$. Entonces, la ecuación de la recta de mínimos cuadrados se expresaría como (véase el problema 13.15):

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x \quad \text{o} \quad y = \left(\frac{\sum xY}{\sum x^2}\right) x \tag{20}$$

En particular, si X es tal que $\sum X = 0$ (es decir, $\bar{X} = 0$); esto se convierte en

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X \tag{21}$$

La ecuación (*20*) implica que $y = 0$ cuando $x = 0$; por lo tanto, la recta de mínimos cuadrados pasa por el punto $(\bar{X}, \bar{Y})$, denominado *centroide* o *centro de gravedad* de los datos.

Si la variable X se toma como la variable dependiente en lugar de la independiente, la ecuación (*17*) se expresa como $X = b_0 + b_1 Y$. Entonces, los resultados son válidos si X y Y se intercambian y si a_0 y a_1 se sustituyen por b_0 y b_1, respectivamente. Sin embargo, la recta de mínimos cuadrados resultante, por lo general, no es igual a la obtenida [véanse los problemas 13.11 y 13.15*d*)].

RELACIONES NO LINEALES

En ocasiones las relaciones no lineales pueden reducirse a relaciones lineales por medio de una trasformación adecuada de las variables (véase el problema 13.21).

LA PARÁBOLA DE MÍNIMOS CUADRADOS

La parábola de mínimos cuadrados que se aproxima al conjunto de puntos (X_1,Y_1), (X_2,Y_2), ..., (X_N,Y_N) tiene por ecuación

$$Y = a_0 + a_1 X + a_2 X^2 \tag{22}$$

donde las constantes a_0, a_1 y a_2 se determinan resolviendo simultáneamente las ecuaciones

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{23}$$

denominadas *ecuaciones normales de la parábola de mínimos cuadrados* (*22*).

Las ecuaciones (*23*) se recordarán fácilmente si se observa que pueden obtenerse multiplicando la ecuación (*22*) por 1, X y X^2, en ese orden, y sumando en ambos lados de las ecuaciones resultantes. Esta técnica suele extenderse para lograr ecuaciones normales de curvas cúbicas de mínimos cuadrados, curvas cuárticas de mínimos cuadrados y, en general, cualquiera de las curvas de mínimos cuadrados correspondientes a la ecuación (*5*).

Al igual que en el caso de la recta de mínimos cuadrados, se pueden simplificar las ecuaciones (*23*) si se elige X de modo que $\sum X = 0$. La simplificación también se da escogiendo las nuevas variables $x = X - \bar{X}$ y $y = Y - \bar{Y}$.

REGRESIÓN

Con frecuencia, basados en datos muestrales, se busca estimar el valor de una variable Y correspondiente a un valor dado de una variable X. Esto se puede lograr estimando el valor de Y a partir de una curva de mínimos cuadrados que se ajuste a los datos muestrales. La curva resultante se llama *curva de regresión de y sobre X*, ya que Y se estima a partir de X.

Si se deseara estimar el valor de X a partir de un valor dado de Y, se utilizaría una *curva de regresión de X sobre Y*, que es igual que intercambiar las variables en el diagrama de dispersión, de tal modo que X sea la variable dependiente y Y la variable independiente. Esto es equivalente a sustituir las desviaciones verticales en la definición de la curva de mínimos cuadrados de la página 287 con desviaciones horizontales.

En general, la recta o curva de regresión de Y sobre X no es igual que la recta o curva de regresión de X sobre Y.

APLICACIONES A SERIES DE TIEMPO

Si la variable independiente X es tiempo, los datos muestran el valor de Y en varios momentos. Los datos ordenados de acuerdo con el tiempo son llamados *series de tiempo*. La recta o curva de regresión de Y sobre X en este caso suele denominarse *recta de tendencia* o *curva de tendencia* y con frecuencia se utiliza con fines de *estimación* o *predicción*.

PROBLEMAS CON MÁS DE DOS VARIABLES

Los problemas que implican más de dos variables a veces son tratados de manera análoga a aquellos con dos variables. Por ejemplo, puede existir una relación entre las tres variables X, Y y Z, que se describe por medio de la ecuación

$$Z = a_0 + a_1X + a_2Y \tag{24}$$

conocido como *ecuación lineal en las variables* X, Y *y* Z.

En un sistema rectangular de coordenadas en tres dimensiones, esta ecuación representa un plano y los puntos muestrales reales $(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots, (X_N, Y_N, Z_N)$ pueden "dispersarse" no demasiado lejos de este plano, denominado *plano de aproximación*.

Por extensión del método de mínimos cuadrados es posible hablar de un *plano de mínimos cuadrados* que se aproxima a los datos. Si se estima Z a partir de valores dados de X y Y, esto se llamaría *plano de regresión de* Z *sobre* X y Y. Las ecuaciones normales, correspondientes al plano de mínimos cuadrados (*24*), están dadas por

$$\begin{aligned}
\sum Z &= a_0N + a_1 \sum X + a_2 \sum Y \\
\sum XZ &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum XY \\
\sum YZ &= a_0 \sum Y + a_1 \sum XY + a_2 \sum Y^2
\end{aligned} \tag{25}$$

y pueden recordarse como obtenidas de la ecuación (*24*), multiplicando por 1, X y Y, sucesivamente, y después sumando.

También pueden considerarse ecuaciones más complicadas que la (*24*). Éstas representan *superficies de regresión*. Si el número de variables excede de 3, se pierde la intuición geométrica, ya que entonces se requiere de espacios de 4, 5,… dimensiones.

Los problemas que implican la estimación de una variable a partir de dos o más variables se denominan problemas de *regresión múltiple* y se estudian con mayor detalle en el capítulo 15.