PART II

MODELOS DE REGRESIÓN LINEAL

Capítulo 2

MODELOS DE REGRESIÓN LINEAL

Un modelo de regresión es cualquier modelo lineal general $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ en la cual $\boldsymbol{X}'\boldsymbol{X}$ es no singular. $\boldsymbol{X}'\boldsymbol{X}$ es no singular si y sólo si la matriz $\boldsymbol{X}_{n\times p}$ tiene rango p. En modelos de regresión, el vector de parámetros $\boldsymbol{\beta}$ es estimable. Veamos formalmente su definición.

Definición 2.1 (Modelo de Regresión Lineal) *Un modelo de regresión lineal es un modelo lineal*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

donde $\boldsymbol{y}$ es un vector $n \times 1$ de observaciones, $\boldsymbol{X}$ es una matriz $n \times p$ de rango p de constantes conocidas, $\boldsymbol{\beta}$ es un vector $p \times 1$ de parámetros no observables, y $\boldsymbol{\varepsilon}$ es un vector $n \times 1$ de variables aleatorias no observables.

2.1 Modelo de Regresión Lineal Simple

En esta sección se considera el modelo de regresión lineal simple. El modelo de regresión lineal simple para n observaciones se puede escribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n \tag{2.2}$$

La designación *simple* indica que existe solamente una variable explicatoria x para predecir la respuesta y, y *lineal* significa que el modelo (2.3) es lineal en β_0 y β_1. Asumimos que y_i son variables observables, los x_i son constantes conocidas (lo cual significa que los mismos

valores de $x_1, x_2, ..., x_n$ deben ser usados en muestreos repetidos), las ϵ_i son variables aleatorias no observables, β_0 y β_1 son constantes desconocidas y son los parámetros del modelo.
Para completar el modelo en (2.3), hacemos los siguientes supuestos adicionales

1. $E(\epsilon_i) = 0$ para todo $i = 1, 2, ..., n$, o, equivalentemente, $E(y_i) = \beta_0 + \beta_1 x_i$.
2. $Var(\epsilon_i) = \sigma^2$ para todo $i = 1, 2, ..., n$, o, equivalentemente, $Var(y_i) = \sigma^2$.
3. $Cov(\epsilon_i, \epsilon_j) = 0$ para todo $i \neq j$, o equivalentemente, $Cov(y_i, y_j) = 0$

El supuesto 1 establece que el modelo (2.3) es correcto, implicando que y_i depende solamente de x_i y que toda la otra variación en y_i es aleatoria. El supuesto 2 asegura que la varianza de ϵ_i o y_i no depende de los valores de x_i (este supuesto es conocido como el supuesto de *homocedasticidad, varianza homogénea o varianza constante*). Bajo el supuesto 3, las variables ϵ_i o y_i están descorrelacionadas unas de otras.

2.1.1 Modelo de Regresión Lineal Simple en Términos Matriciales

En esta sección desarrollamos el modelo de regresión lineal simple en términos matriciales. Recordemos que el modelo de regresión lineal simple para n observaciones se puede escribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n \tag{2.3}$$

Esto implica

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned} \tag{2.4}$$

lo cual se puede escribir como

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Haciendo

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{2.5}$$

se tiene que las ecuaciones 2.8 se pueden escribir de manera compacta como

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.6}$$

el cual es la ecuación de un modelo lineal como el estudiado en el capítulo anterior.

Note que la columna de unos en la matriz $\boldsymbol{X}$ puede verse como consistente de la constante $x_0 = 1$ en el modelo de regresión alternativo:

$$y_i = beta_0 x_0 + \beta_1 x_i + \epsilon_i \quad \text{donde } x_0 = 1$$

Por lo tanto, la matriz $\boldsymbol{X}$ puede ser considerada a contener un vector columna de unos y otro vector columna que contiene las observaciones x_i de la variable predictora.

En cuanto a los supuestos, el supuesto 1 el cual establece que $E(\epsilon_i) = 0$ para $i = 1, 2, ..., n$ se puede escribir como $E(\boldsymbol{\varepsilon}) = 0$. Y los supuestos 2 y 3, $Var(\epsilon_1) = \sigma^2$ y $Cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$, se pueden resumir como $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ ya que

$$Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

EJEMPLO 2.1 Tomado de Rencher 2008

Estudiantes en una clase de estadística expusieron que hacer la tarea no ayudaba a prepararlos para el examen a mitad del período. El puntaje y del examen y el puntaje x de la tarea (promediados hasta la mitad del período) para los 18 estudiantes en la clase fueron los siguientes

y	x	y	x	y	x
95	96	72	89	35	0
80	77	66	47	50	30
0	0	98	90	72	59
0	0	90	93	55	77
79	78	0	18	75	74
77	64	95	86	66	67

Vamos e representar matricialmente las observaciones de los primeros 6 estudiantes. Por lo tanto se tiene

$$\boldsymbol{y} = \begin{pmatrix} 95 \\ 80 \\ 0 \\ 0 \\ 79 \\ 77 \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 96 \\ 1 & 77 \\ 1 & 0 \\ 1 & 0 \\ 1 & 78 \\ 1 & 64 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{pmatrix}$$

□

2.1.2 Principio Básico del Modelo de Regresión Lineal Simple

Como en este caso sobre la variable respuesta y solo "influye" una variable predictora X, se quiere determinar la ecuación de una recta (linealidad) la cual se ajuste lo más posible a los datos. Un diagrama de dispersión ofrece una idea bastante aproximada sobre la existencia o no de una relación lineal entre dos variables, además puede utilizarse como una forma de cuantificar el grado de relación lineal existente entre dos variable: basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.

Ahora bien, aunque un diagrama de dispersión permite formarse una primera impresión muy rápida sobre el tipo de relación existente entre dos variables, utilizarlo como una forma de cuantificar esa relación tiene un serio inconveniente: la relación entre dos variables no siempre es perfecta o nula; de hecho, habitualmente no es ni lo uno ni lo otro.

Supongamos que disponemos de un pequeño conjunto de datos como los que se muestran en la tabla 2.1

Table 2.1 Datos

x_1	y_1	x_3	y_8
x_1	y_2	x_3	y_9
x_2	y_3	x_4	y_{10}
x_2	y_4	x_5	y_{11}
x_2	y_5	x_5	y_{12}
x_3	y_6	x_5	y_{13}
x_3	y_7		

Primero observemos como para el mismo valor de x podemos obtener valores distintos de y. Es decir que una vez fijado el valor de x digamos en x_1, se observaron tres valores de la variable respuesta, y_1, y_2, y_3. Como se comentó anteriormente un buen punto de partida para formarnos una primera impresión de esa relación podría ser la representación de la nube de puntos, tal como muestra el diagrama de dispersión de la figura 2.1.

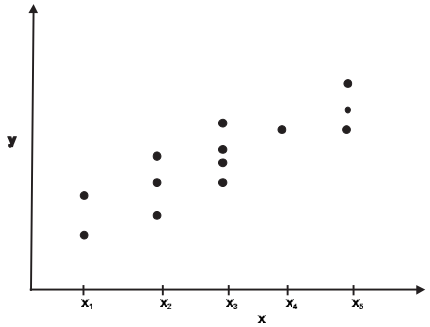

Figure 2.1 Nube de puntos

De la figura podemos observar que además de verse un comportamiento aproximadamente lineal entre los puntos, se nota que la dispersión de los datos para cada valor de x es muy parecida, el cual cual es el supuesto mencionado en la definición del modelo (ver figura 2.2).

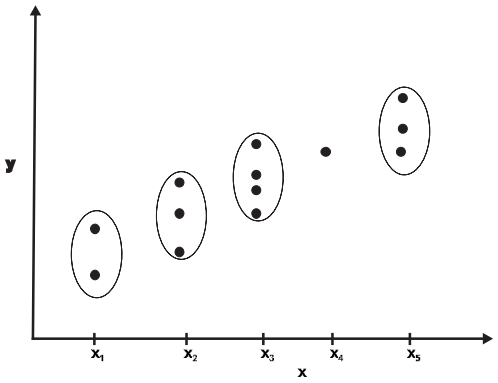

Figure 2.2 Nube de puntos

Ahora bien, para obtener la recta que mejor se ajuste a los datos, dicha recta debería pasar por el promedio de las observaciones para cada valor de x, como se muestra en la gráfica de la figura 2.3. En el próximo capítulo veremos como calcular dicha recta.

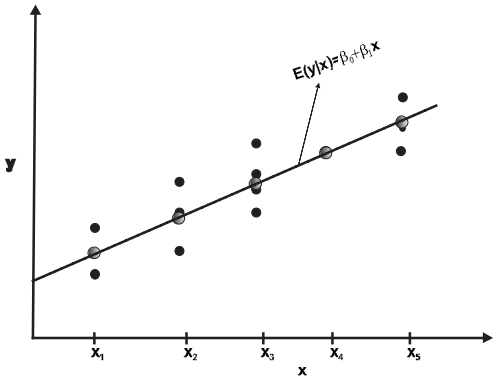

Figure 2.3 Nube de puntos y recta de regresión lineal

En la siguiente figura se muestra el gráfico de dispersion de los datos del ejemplo 2.1. En dicho gráfico se aprecia una aparente relación lineal entre las variables.
Podríamos trazar diversas rectas que pasen cerca de un gran numero de puntos como se observa en la figura 2.5, pero la idea es conseguir aquella que se ajuste mejor a los datos. El cálculo de dicha recta será nuestro objetivo en el siguiente capítulo.

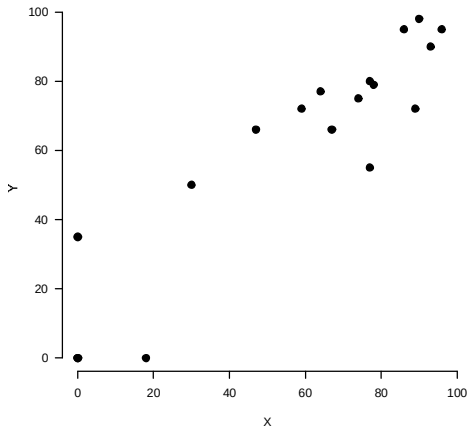

Figure 2.4 Diagrama de dispersion para los datos de tarea y puntos en el examen

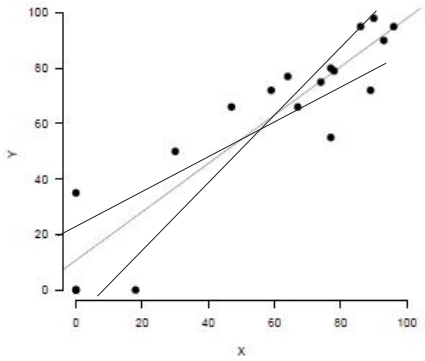

Figure 2.5 Múltiples rectas para los datos de tarea y puntos en el examen

2.2 Modelo de Regresión Lineal Múltiple

En regresión múltiple, intentamos predecir una variable dependiente o respuesta y sobre la base de una relación lineal asumida con varias variables independientes o predictoras $x_1, x_2, ..., x_n$. Además de construir un modelo para la predicción, se establecen algunas medidas o técnicas para medir el ajuste del modelo creado.

2.2.1 El Modelo

El modelo de regresión lineal múltiple para k variables predictoras y n observaciones puede ser expresado como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,\ i = 1, 2, ..., n \tag{2.7}$$

La designación *múltiple* indica que existen varias variable explicatorias x_j, $j = 1, ..., k$ para predecir la respuesta y, y *lineal* significa que el modelo (2.7) es lineal en $\beta_0, \beta_1, ..., \beta_k$. Asumimos que y_i son variables observables, las x_{ij} son constantes conocidas, las ϵ_i son variables aleatorias no observables, $\beta_0, \beta_1, ..., \beta_k$ son constantes desconocidas y son los parámetros del modelo.
Al igual que para el modelo de regresión lineal simple, se tienen los siguientes supuestos adicionales

1. $E(\epsilon_i) = 0$ para todo $i = 1, 2, ..., n$, o, equivalentemente, $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$.
2. $Var(\epsilon_i) = \sigma^2$ para todo $i = 1, 2, ..., n$, o, equivalentemente, $Var(y_i) = \sigma^2$.
3. $Cov(\epsilon_i, \epsilon_j) = 0$ para todo $i \neq j$, o equivalentemente, $Cov(y_i, y_j) = 0$

2.2.2 Modelo de Regresión Lineal Simple en Términos Matriciales

En esta sección desarrollamos el modelo de regresión lineal múltiple en términos matriciales. Recordemos que el modelo de regresión lineal múltiple para n observaciones se puede escribir como

$$y_i = \beta_0 + beta_1 x_{i1} + beta_2 x_{i2} + \cdots + beta_k x_{ik} + \epsilon_i,\ i = 1, 2, ..., n$$

Esto implica

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n \end{aligned} \tag{2.8}$$

Estas n ecuaciones se pueden escribir en forma de matriz como

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{pmatrix}$$

o

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.9}$$

Los tres supuestos sobre ϵ_i se pueden expresar en términos del modelo en 2.9:

1. $E(\boldsymbol{\varepsilon}) = 0$ ó $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$.

2. $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ ó $Cov(\boldsymbol{y}) = \sigma^2\mathbf{I}$

Note que el supuesto $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ incluye ambos supuestos $Var(\epsilon_i) = \sigma^2$ y $Cova(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$.
La matriz $\boldsymbol{X}$ en 2.9 es $n \times (k+1)$. En este curso asumimos que $n > k+1$ y que $rang(\boldsymbol{X}) = k+1$. Si $n < k+1$ o si existe relación lineal entre las x, poe ejemplo, $x_5 = \sum_{j=1}^{4} x_j/4$, entonces $\boldsymbol{X}$ no será de rango completo por columnas.

EJEMPLO 2.2

Suponga que se tienen los siguientes datos [Ver Freund y Minton (1979, pp- 36 -39)] de la tabla 2.2

Table 2.2 Datos del ejemplo

Observación	y	x_1	x_2
1	2	0	2
2	3	2	6
3	2	2	7
4	7	2	5
5	6	4	9
6	8	4	8
7	10	4	7
8	7	6	10
9	8	6	11
10	12	6	9
11	11	8	15
12	14	8	13

Vamos e representar matricialmente las observaciones. Por lo tanto se tiene

$$\boldsymbol{y} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ \vdots \\ 14 \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ \vdots & \vdots & \vdots \\ 1 & 8 & 13 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_12 \end{pmatrix}$$

□

Los parámetros $\boldsymbol{\beta}$ en 2.9 son llamados los *coeficientes de regresión.* Para enfatizar su efecto colectivo, ellos algunas veces son referidos como *coeficientes de regresión parcial.* La palabra parcial tiene un significado tanto matemático como estadístico. Matemáticamente, la derivada parcial de $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ con respecto a x_1, por ejemplo, es β_1. Así β_1 indica el cambio en $E(y)$ a consecuencia de un incremento de una unidad en x_1 cuando $x_2, x_3, ..., x_k$ se mantienen constantes. Estadísticamente, β_1 muestra el efecto de x_1 sobre $E(y)$ en la presencia de las otras x. Este efecto debería típicamente ser

diferente del efecto de x_1 sobre $E(y)$ si las otras x no estuviesen presentes en el modelo. Así, por ejemplo, β_0 y β_1 en

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

será usualmente diferente de β_0^* y β_1^* en

$$y = \beta_0^* + \beta_1^* x_1 + \epsilon$$

Sólo si x_1 y x_2 son ortogonales entonces $\beta_0 = \beta_0^*$ y $\beta_1 = \beta_1^*$. El cambio en los parámetros cuando un x es borrado del modelo se ilustra (con estimaciones) en el siguiente ejemplo

EJEMPLO 2.3

Tomando los datos de la tabla 2.2 ejemplo anterior, se estimaron tres modelos, los dos primeros tomando en cuanta solo una variable independiente en cada caso, y el último tomando en cuenta la dos variable. Los modelos estimados son los siguientes

$$\begin{aligned} \hat{y} &= 1.86 + 1.30x_1, \\ \hat{y} &= 0.86 + 0.78x_2, \\ \hat{y} &= 5.37 + 3.01x_1 - 1.29x_2. \end{aligned}$$

Como se esperaba, los coeficientes cambian del modelo con todas las variables (modelo completo) a los modelos con una sola variable (modelo reducido). Incluso note que el signo del coeficiente de x_2 cambia de 0.78 a -1.29.

Los valores de y y x_2 son graficados en la figura 2.6 junto con la ecuación de predicción $\hat{y} = 0.86 + 0.78x_2$. La tendencia lineal es evidente.

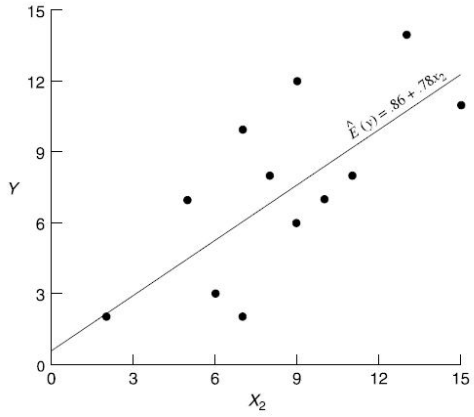

Figure 2.6 Regresión de y sobre x_2 ignorando x_1

En la figura 2.7 se muestra el gráfico como en la figura 2.6, excepto que cada punto es etiquetado con el valor de x_1. Examinando los valores de y sobre x_2 para valores fijos de x_1 (2,4,6, o 8) muestra una pendiente negativa para la relación. Esta relación negativa son mostradas como regresiones parciales de y sobre x_2 para cada valor de

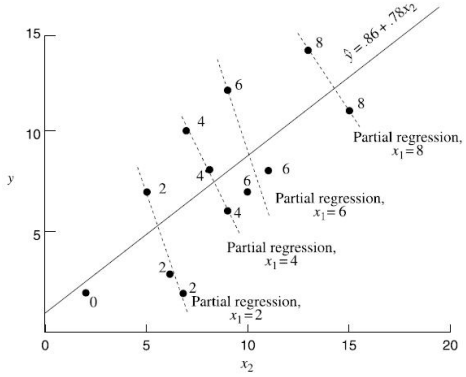

Figure 2.7 Regresión de y sobre x_2 mostrando el valor de x_1 en cada punto y regresiones parciales de y sobre x_2

x_1. El coeficiente de regresión parcial $\hat{\beta}_2 = -1.29$ refleja las pendiente negativas de estas cuatro regresiones parciales.

PART III

DISEÑO