# Modelos Lineales

Rubén Martín

# Índice general

# Capítulo 1

# Introducción a los Modelos Lineales y a los Modelos de Regresión

## 1.1. Definición del Modelo Lineal

Sea $Y$ una variable de la cual se desea estudiar algunos aspectos sobre su comportamiento (predecir valores futuros, comprobar si se comporta de igual manera ante influencias externas diferentes,...)

Para ese análisis consideremos un conjunto de variables *auxiliares*$(X_1, ..., X_k)$que se cree pueden aportar información acerca del problema concreto que se desea estudiar.

Dada la naturaleza de ambos tipos de variables es usual denominar variable *explicada* o *dependiente* a la variable objeto de estudio, $Y$, mientras que a las otras variables se les conoce como variables *explicativas* o *independientes*.

Los modelos lineales adoptan tal denominación debido a que el modelo para estudiar el comportamiento de la variable dependiente vía las independientes es de la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$$

Para realizar el análisis que pretendemos será necesario disponer de una muestra formada por $N$, $(N > k+1)$, observaciones de dichas variables$(y_i; x_{i1}, ..., x_{ik})$, $i = 1, ..., N$, las cuales verificarán las expresiones

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

que matricialmente se pueden expresar en la forma

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & \cdots & x_{1k} \\ 1 & x_{11} & x_{12} & \cdots & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & \cdots & x_{Nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

y en la fórmula reducida $y = X\beta + \epsilon$, donde $X$ es la llamada *Matriz de Diseño*, cuyas columnas contienen las observaciones de las variables independientes más una columna de unos(si incluimos término independiente). A esta expresión genérica se le conoce como *Modelo Lineal General*.

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

En este planteamiento general quedan multitud de cuestiones por precisar como pueden ser las siguientes:

- La distribución de la variable de perturbación $\epsilon$
- La aleatoriedad o no de los parámetros $\beta_j$
- La aleatoriedad o no de las variables explicativas
- Tipo de matriz de diseño así como su rango

Así se puede hacer una clasificación de los modelos lineales atendiendo a la naturaleza y las interrelaciones de los elementos del modelo. También se pueden clasificar según la dimensión de las variables, distinguiéndose así entre el Modelo Lineal General Univariante y Multivariante. Por último, se dirá que el modelo es simple si sólo considera una variable explicativa, mientras que diremos que es múltiple si existen varias.

## 1.2. Algunos Tipos de Modelos Lineales

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

$$y = X\beta + \epsilon$$

- Se prefijan valores observados $x_1, ..., x_k$ para $X_1, ...X_k$
- $Y$ observada fijados valores $x_1, ..., x_k$ de las variables independientes
- $\beta_i$ no aleatorios
- $\epsilon_i$ aleatorios

MODELOS DE REGRESIÓN CONDICIONADA
*Modelo de Regresión Lineal*

- $X_1, ...X_k$ con valores 0 ó 1 que determinan condiciones del fenómeno
- $Y$ observada para valores de las condiciones anteriores
- $\beta_i$ no aleatorios
- $\epsilon_i$ aleatorios

MODELOS DE RELACIÓN FUNCIONAL
*Modelos ANOVA*

- $X_1, ..., X_k$ con valores 0 ó 1 y valores $x_1, ..., x_k$
- $Y$ observada para valores $x_1, ..., x_k$
- $\beta_i$ no aleatorios
- $\epsilon_i$ aleatorios

MODELOS DE LA RELACIÓN FUNCIONAL
*Diseños experimentales tipo ANCOVA*

- $X_1, ..., X_k$ con valores 0 ó 1 que determinan condiciones del fenómeno
- $Y$ observada para dichas condiciones
- $\beta_i$ no aleatorios, independientes de los términos $\epsilon_i$
- $\epsilon_i$ aleatorios

MODELOS DE LA RELACIÓN FUNCIONAL
*Diseños experimentales de Efectos Aleatorios (Componentes de la Varianza*

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & \cdots & x_{1k} \\ 1 & x_{11} & x_{12} & \cdots & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & \cdots & x_{Nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}, (N > k+1)$$

$$\downarrow$$

$$y = X\beta + \epsilon$$

$$\downarrow$$

MODELO LINEAL DE GAUSS MARKOV

- $\epsilon_i$ aleatorias con media cero, varianza $\sigma^2$ e incorreladas
- $\beta_i$ no aleatorios
- Los valores de $X$ están prefijados

## 1.3. Modelo Lineal de Gauss-Markov

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & \cdots & x_{1k} \\ 1 & x_{11} & x_{12} & \cdots & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & \cdots & x_{Nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}, (N > k+1)$$

$$\downarrow$$

$$y = X\beta + \epsilon$$

$$\downarrow$$

MODELO LINEAL DE GAUSS MARKOV

- $E[\epsilon] = 0, Cov[\epsilon] = \sigma^2 \mathbf{I}_N$
- $\beta$ no aleatorio
- $X$ no aleatoria
  - $Rag(X) = k+1$. Modelo de rango completo
  - $Rag(X) < k+1$. Modelo de rango no completo

INFERENCIA

- Estimación:
  - Mínimos cuadrados
  - Máxima verosimilitud(supuesta normalidad en los errores)
- Contraste

GENERALIZACIÓN
Modelo de Aitken: $Cov(\varepsilon) = \sigma^2 V,\ V$ *conocida*

## 1.4. Ejemplos de Modelos de Gauss-Markov

PROBLEMA DE UNA MUESTRA

$$y \to N_1[\mu;\sigma^2] \Rightarrow y = \mu + \epsilon \ (\epsilon \to N_1[0;\sigma^2])$$

$$\downarrow$$

$$\{y_1, ..., y_N\} \ m.a.s.$$

$$\downarrow$$

$$y_i = \mu + \epsilon_i \ (\epsilon_i \to N_1[0;\sigma^2])$$

$$\downarrow$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

$$\downarrow$$

$$y = X\mu + \epsilon$$

PROBLEMA DE DOS MUESTRAS(rango completo)

$$y_1 \to N_1[\mu_1;\sigma^2];\ y_2 \to N_1[\mu_2;\sigma^2], \text{ independientes}$$

$$\downarrow$$

$$\{y_{11}, ..., y_{1N_1}\},\ \{y_{21}, ..., y_{2N_2}\} \ m.a.s.$$

$$\downarrow$$

$$y_{ij} = \mu_i + \epsilon_{ij} \ (\epsilon_{ij} \to N_1[0;\sigma^2])$$

$$\downarrow$$

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1N_1} \\ y_{21} \\ \vdots \\ y_{2N_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{pmatrix}$$

$$\downarrow$$

$$y = X\beta + \epsilon$$

PROBLEMA DE DOS MUESTRAS(rango no completo)

$$y_1 \to N_1[\mu_1;\sigma^2];\ y_2 \to N_1[\mu_2;\sigma^2], \text{ independientes}$$

$$\downarrow$$

$$\{y_{11}, ..., y_{1N_1}\},\ \{y_{21}, ..., y_{2N_2}\} \ m.a.s.$$

$$\downarrow$$

$$y_{ij} = \mu_i + \alpha_i + \epsilon_{ij} \ (\epsilon_{ij} \to N_1[0;\sigma^2])$$

$$\downarrow$$

$$\left(\begin{array}{c} y_{11} \\ \vdots \\ \hline y_{1N_1} \\ y_{21} \\ \vdots \\ y_{2N_2} \end{array}\right) = \left(\begin{array}{ccc} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ \hline 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{array}\right) \left(\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \end{array}\right) + \left(\begin{array}{c} \epsilon_{11} \\ \vdots \\ \hline \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2N_2} \end{array}\right)$$

$$\downarrow$$

$$y = X\beta + \epsilon$$

PROBLEMA DE K MUESTRAS O ANOVA DE UNA VÍA(rango completo)

$$y_1 \to N_1[\mu_1; \sigma^2]; \cdots ;\ y_i \to N_1[\mu_i; \sigma^2], \text{ independientes}$$

$$\downarrow$$

$$\{y_{11}, ..., y_{1N_1}\}, \cdots, \{y_{i1}, ..., y_{iN_i}; \cdots ; \{y_{k1}, ..., y_{kN_k}\}\ m.a.s.$$

$$\downarrow$$

$$y_{ij} = \mu_i + \epsilon_{ij}\ (\epsilon_{ij} \to N_1[0; \sigma^2])$$

$$\downarrow$$

$$\left(\begin{array}{c} y_{11} \\ \vdots \\ \hline y_{1N_1} \\ y_{21} \\ \vdots \\ \hline y_{2N_2} \\ y_{k1} \\ \vdots \\ y_{kN_k} \end{array}\right) = \left(\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \hline 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{array}\right) \left(\begin{array}{c} \mu_1 \\ \mu_2 \\ \vdots \\ \alpha_k \end{array}\right) + \left(\begin{array}{c} \epsilon_{11} \\ \vdots \\ \hline \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \hline \epsilon_{2N_2} \\ \epsilon_{k1} \\ \vdots \\ \epsilon_{kN_k} \end{array}\right)$$

$$\downarrow$$

$$y = X\beta + \epsilon$$

PROBLEMA DE K MUESTRAS O ANOVA DE UNA VÍA(rango no completo)

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}\ (\epsilon_{ij} \to N_1[0; \sigma^2])$$

$$\downarrow$$

$$\left(\begin{array}{c} y_{11} \\ \vdots \\ \hline y_{1N_1} \\ y_{21} \\ \vdots \\ \hline y_{2N_2} \\ y_{k1} \\ \vdots \\ y_{kN_k} \end{array}\right) = \left(\begin{array}{ccccc} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{array}\right) \left(\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{array}\right) + \left(\begin{array}{c} \epsilon_{11} \\ \vdots \\ \hline \epsilon_{1N_1} \\ \epsilon_{21} \\ \vdots \\ \hline \epsilon_{2N_2} \\ \epsilon_{k1} \\ \vdots \\ \epsilon_{kN_k} \end{array}\right)$$

$$\downarrow$$

$$y = X\beta + \epsilon$$

## 1.5. Regresión

**Regresión:** Búsqueda de una función que exprese la relación dos o más variables.

**Variables:** Dependiente o explicada($Y$). Explicativas, independientes o regresores($X_1, ..., X_k$)

**Orígenes:**

- **Astronomía y Física:** Laplace y Gauss
- **Biología:** Galton(acuñó el término regresión)

**Formulación del modelo:**

- Encontrar $g$ tal que $Y = g(X_1, ..., X_k; \epsilon)$
- **¿Quién es $g$?** Distintos tipos de regresión
  - **Regresión lineal:** $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon$

# Capítulo 2

# El Modelo de Regresión Lineal Simple Univariante (I)

## 2.1. Hipótesis Básicas del Modelo

Sea $Y$ una variable que representa una característica de una población, característica objeto de estudio y sobre la cual se desea conocer diversos aspectos de su comportamiento. Para ello disponemos de la información suministrada por otra variable $X$, cuyos valores pueden ser determinados *a priori*. A $Y$ la conoceremos como la variables dependiente (o variable explicada o regresando), mientras que $X$ es la variable independiente (o variable explicativa o regresor).

Admitiremos que la hipótesis estructural básica del modelo es

$$Y = \beta_0 + \beta_1 X + \epsilon$$

o sea, la relación entre ambas variables es de tipo lineal en los parámetros del modelo. En este modelo supondremos:

- $X$ es una variable cuyos valores son conocidos al observar los valores de $Y$
- $\epsilon$ es una variable aleatoria que engloba un conjunto de factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud pero que de forma conjunta debe tenerse en cuenta en la especificación y tratamiento del modelo.
- $\beta_0$ y $\beta_1$ son constantes fijas(no aleatorias) pero desconocidas, cuyos valores deberán ser estimados.

Como hemos dicho anteriormente, para cada valor xi fijo de la variable independiente (condición experimental) se dispondrá de una realización de la variable dependiente. Por lo tanto tendremos una muestra de pares de valores $(x_i, y_i)$, $i = 1, ..., N$ $(N > 2)$.

Dada la estructura funcional impuesta del modelo, para cada valor $x_i$ fijo se verifica

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ ; \ i = 1, ..., N$$

donde las variables $\epsilon_i$ (perturbaciones) se consideran realizaciones de la variable de error $\epsilon$.

Notemos que en lo que hasta ahora se ha dicho ya hay una serie de hipótesis establecidas. Además de ellas, se añaden las siguientes hipótesis sobre las variables de perturbación y la variable explicativa o independiente:

- Las perturbaciones tienen media cero

$$E[\epsilon_i] = 0 \ ; \ i = 1, ..., N$$

- Las perturbaciones tienen varianza constante(hipótesis de homocedasticidad)

$$Var[\epsilon_i] = \sigma^2 \ ; \ i = 1, ..., N$$

- Las perturbaciones son incorreladas entre sí(hipótesis de incorrelación)

$$Cov[\epsilon_i, \epsilon_j] = E[\epsilon_i \epsilon_j] = 0 \ ; \ i, j = 1, ..., N \ (i \neq j)$$

- Los valores de la variable $X$ no son todos iguales, o sea, al menos hay dos observaciones distintas o, lo que es lo mismo, la variable $X$ es no degenerada.

**Nota 1.1.** *Observemos que la formulación del modelo, junto con las hipótesis establecidas nos conduce a un modelo lineal de Gauss-Markov con matriz de diseño*

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & x_N \end{pmatrix}$$

*Puesto que la variable explicativa es no degenerada, las columnas de dicha matriz no pueden ser proporcionales y con ello su rango es 2. Por lo tanto el modelo es de rango completo.*

Las hipótesis que atañen a las variables perturbación pueden ser formuladas en términos de la variable explicada puesto que del hecho de que la variable explicativa sea no aleatoria (ni los efectos tampoco) se desprende que toda la carga aleatoria del modelo descansa sobre las variables de perturbación y por lo tanto la variable Y retoma el carácter aleatorio de ellas. Así se pueden expresar las tres primeras hipótesis anteriores en la forma siguiente:

- La esperanza de la respuesta depende linealmente de $X$: $E[y_i] = \beta_0 + \beta_1 x_i \ ; \ i = 1, ..., N$
  Realmente deberíamos escribir $E[y_i|X = x_i] = \beta_0 + \beta_1 x_i \ ; \ i = 1, ..., N$
- La varianza de las variables $y_i$ es constante: $Var[y_i] = \sigma^2 \ ; \ i = 1, ..., N$
- Las observaciones $y_i$ son incorreladas entre sí: $Cov[y_i, y_j] = 0 \ ; \ i, j = 1, ..., N \ (i \neq j)$

**Nota 1.2.** *$\beta_0$ representa el valor medio de la variable $Y$ cuando la variable $X$ vale cero. Asimismo $\beta_1$ es el incremento que experimenta la media de $Y$ cuando $X$ aumenta en una unidad.*

El modelo incluye otra hipótesis, si bien no es preciso para todo lo que se va a realizar sobre él. En concreto no hará falta en lo que concierne a la estimación del modelo por el método de mínimos cuadrados (si bien sí la hará cuando la estimación se realiza por máxima verosimilitud), aunque resultará imprescindible en el momento en que sean necesarias las distribuciones de los estadísticos involucrados en el proceso para establecer contrastes de hipótesis e intervalos de confianza. Esta hipótesis es la siguiente:

- Las variables de perturbación son independientes y están igualmente distribuidas según una ley normal de media 0 y varianza $\sigma^2$

Asimismo se puede reformular esta hipótesis refiriéndola a las variables $y_i$:

- La distribución de $y_i$, para cada $x_i$, es normal de media $\beta_0 + \beta_1 x_i$ y varianza $\sigma^2$, siendo todas las distribuciones independientes

### 2.1.1. Comentarios a las Hipótesis del Modelo

1. La hipótesis principal del modelo es que la media de la distribución de $Y$, para cada valor de $X$ fijo, varía de forma lineal con dicho valor. Esta hipótesis, en la medida que se pueda, debe ser comprobada siempre ya que condiciona toda la construcción del modelo. En cualquier caso conviene tener en cuenta que una relación lineal debe considerarse en general como una aproximación simple, en un rango limitado, a una relación más compleja. En consecuencia, es necesario tener presente el rango de valores dentro del cual se va a trabajar y el peligro de extrapolar la relación fuera de ese rango.

2. La hipótesis de homocedasticidad no se cumplirá si la variabilidad depende, por ejemplo, de las observaciones de la variable independiente. Por ejemplo, si se pretende estudiar el ahorro en función de la renta en varias familias, podemos fácilmente pensar que la variabilidad del ahorro dependerá, de forma fuerte, del nivel de renta de las familias ya que, a renta superior, una familia tiene una mayor flexibilidad a la hora de qué hacer con su dinero, teniendo la posibilidad de ahorrar o consumir, mientras que las familias de renta inferior tendrán menos posibilidades de ahorrar, moviéndose por lo tanto en una franja más estrecha y menos flexible.

3. La incorrelación entre las perturbaciones es esperable en situaciones estáticas, o sea, cuando las observaciones correspondan al mismo periodo temporal, pero no lo será tanto en situaciones dinámicas en las que se mide la variable respuesta a lo largo del tiempo. Por ejemplo, si pretendemos estudiar dos variables de

índole económica en distintos países durante un mismo año, como pueden ser el producto interior bruto como variable independiente y el consumo como variable dependiente, es de suponer, en principio, que no tiene por qué haber una dependencia entre las observaciones. Sin embargo ese mismo tipo de estudio hecho en un país concreto a lo largo de varios años puede llevar implícito el hecho de que los factores recogidos en la perturbación hayan evolucionado en el tiempo y, por lo tanto, haya algún tipo de correlación a lo largo del tiempo entre las perturbaciones.

## 2.2. Estimación del Modelo por Mínimos Cuadrados Ordinarios

### 2.2.1. Obtención de Estimadores

Consideremos el modelo de regresión lineal simple $Y = \beta_0 + \beta_1 X + \epsilon$

Nuestra intención es encontrar los valores de $\beta_0$ y $\beta_1$ tales que expliquen de la mejor forma posible la relación de tipo lineal que liga a las variables en estudio. Para ello tendremos que buscar dos estimaciones de dichos parámetros, llamémoslas $\hat{\beta}_0$ y $\hat{\beta}_1$

Bajo este supuesto, y para cada valor de $x$ de $X$, la predicción que sobre la variable $Y$ se haría es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

expresión que genera, al variar $X$ en su rango, una recta, llamada *recta de regresión.*

Para realizar el proceso de estimación necesitamos una muestra de la variable dependiente. Cada elemento de dicha muestra se obtiene fijado un valor $x_i$ de $X$. De esta forma tenemos las relaciones

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ ; \ i = 1, ..., N, \ (N > 2)$$

A continuación debemos fijar un criterio para la estimación. En primer lugar emplearemos el criterio de mínimos cuadrados ordinarios, según el cual hemos de minimizar, en $\beta_0$ y $\beta_1$, la suma de los cuadrados de los errores. Es decir, el problema que hay que resolver es

$$\underset{\beta_0,\beta_1}{\text{Min}} \sum_{i=1}^{N} \epsilon_i^2$$

Intuitivamente podemos observar que lo que se hace es minimizar el efecto de las perturbaciones de forma global(piénsese en el hipotético caso en que esa suma valiera cero).

Matemáticamente hablando el problema se traduce en minimizar la función de dos variables

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

Para la resolución técnica se deriva la función anterior respecto de los parámetros y se plantea el sistema de ecuaciones

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0, \qquad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

resultando así

$$-2 \sum_{i=1}^{N} (y_i \beta_0 - \beta_1 x_i) = 0, \qquad -2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

o lo que es lo mismo

$$\sum_{i=1}^{N} \epsilon_i = 0, \qquad \sum_{i=1}^{N} \epsilon_i x_i = 0$$

que da origen al denominado sistema de ecuaciones normales

$$\sum_{i=1}^{N} y_i = N\beta_0 + \beta_1 \sum_{i=1}^{N} x_i$$

$$\sum_{i=1}^{N} y_i x_i = \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2$$

Llamando $\hat{\beta}_0$ y $\hat{\beta}_1$ a la solución del sistema, esta es:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}$$

Además, la matriz hessiana, en el punto $(\hat{\beta}_0, \hat{\beta}_1)$, es

$$H(\hat{\beta}_0, \hat{\beta}_1) = \begin{pmatrix} 2N & 2N\bar{x} \\ 2N\bar{x} & 2\sum_{i=1}^N x_i^2 \end{pmatrix}$$

matriz que claramente es definida positiva, gracias a la hipótesis establecida sobre la variable explicativa, puesto que su determinante es $4N^2 s_x^2$.

EJERCICIO: Plantear, resolver el sistema de ecuaciones normales y comprobar que la solución minimiza la función objetivo.

De esta forma la recta de regresión estimada queda en la forma

$$\hat{y} - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Definimos ahora los residuos mínimo-cuadráticos como la diferencia entre el valor real observado de la variable dependiente y el predicho por la recta de regresión, es decir

$$e_i = y_i - \hat{y}_i = (y_i - \bar{y}) - \frac{s_{xy}}{s_x^2}(x_i - \bar{x})$$

con lo cual se puede concluir que el valor mínimo que alcanza la función $S(\beta_0, \beta_1)$ es $\sum_{i=1}^N e_i^2$. De la anterior expresión para los residuos se puede deducir lo siguiente:

$$\sum_{i=1}^N e_i = 0, \quad \bar{\hat{y}} = \bar{y}, \qquad \sum_{i=1}^N x_i e_i = 0, \qquad \sum_{i=1}^N (\hat{y} - \bar{y}) e_i = 0$$

EJERCICIO: Verificar estas igualdades partiendo de la expresión anterior para $e_i$.

De dichas relaciones se puede concluir que $y_i = \hat{y}_i + e_i$, donde $\hat{y}_i$ es la parte estimada y $e_i$ es la parte residual debida a la regresión, siendo ambas incorreladas.

EJERCICIO: Realizar la estimación m.c. en el modelo sin término constante: $Y = \beta_1 X + \varepsilon$.

## 2.3. COMPLEMENTO: Caso de Datos Repetidos

En muchas ocasiones la experiencia que se pretende estudiar se diseña de forma que, para cada valor fijado de la variable explicativa, se observan diversos valores de la variable dependiente. Nosotros nos referiremos a esta situación con el nombre de *datos repetidos*.

Si suponemos que son $d$, $d > 2$, los valores distintos de la variable explicativa fijados y que, para cada uno, se observan $n_i$ valores de la variable dependiente, los datos suelen presentarse en la forma