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 ϵ
- La aleatoriedad o no de los parámetros β_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
- β_i no aleatorios
- ϵ_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
- β_i no aleatorios
- ϵ_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 $x_1, ..., x_k$
- β_i no aleatorios
- ϵ_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
- β_i no aleatorios, independientes de los términos ϵ_i
- ϵ_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

- ϵ_i aleatorias con media cero, varianza σ^2 e incorreladas
- β_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

- $E[\epsilon] = 0, Cov[\epsilon] = \sigma^2 \mathbf{I}_N$
- β 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
- ϵ 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.
- β_0 y β_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 ϵ_i (perturbaciones) se consideran realizaciones de la variable de error ϵ.

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. *β_0 representa el valor medio de la variable Y cuando la variable X vale cero. Asimismo β_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 σ^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 σ^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 β_0 y β_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 $\widehat{\beta_0}$ y $\widehat{\beta_1}$

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

$$\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

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 β_0 y β_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 $\widehat{\beta_0}$ y $\widehat{\beta_1}$ a la solución del sistema, esta es:

$$\widehat{\beta_1} = \frac{s_{xy}}{s_x^2}$$

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

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

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

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

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

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

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

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:

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

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

De dichas relaciones se puede concluir que $y_i = \widehat{y}_i + e_i$, donde $\widehat{y}_i$ 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: $Y = \beta_1 X + \varepsilon$.

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

x_1	y_{11}	$\cdots$	$\cdots$	y_{1n_1}
x_2	y_{21}	$\cdots$	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_d	y_{d1}	$\cdots$	$\cdots$	y_{dn_d}

Gráficamente

El modelo se escribe en la forma $y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$, $i = 1, ..., d$; $j = 1, ..., n_i$

Notemos que no estamos tratando con un modelo distinto sino con el mismo anterior que se ha reescrito acorde con la estructura de los datos. No obstante, es interesante desarrollar de nuevo la estimación mínimo-cuadrática en esta forma para asentar mejor la idea de la regresión al comprobar qué es lo que hace el modelo en esta situación.

La función que hay que minimizar se escribe ahora como

$$S(\beta_0,\beta_1)=\sum_{i=1}^{d}\sum_{j=1}^{n_i}(y_{ij}-\beta_0-\beta_1 x_i)^2$$

Razonando igual que antes se llega al sistema de ecuaciones normales

$$\sum_{i=1}^{d} n_i\bar{y}_i = N\beta_0+\beta_1\sum_{i=1}^{d} n_i x_i$$

$$\sum_{i=1}^{d} n_i\bar{y}_i x_i = \beta_0\sum_{i=1}^{d} n_i x_i+\beta_1\sum_{i=1}^{d} n_i x_i^2$$

cuya solución, obviamente, es la misma obtenida anteriormente.

En este caso la estimación considera la nube de puntos $\{\mathbf{x_i},\bar{\mathbf{y}}_\mathbf{i}\}_{\mathbf{i=1,...,d}}$, tratando cada punto con la frecuencia $\mathbf{n_i}$ observada.

Asimismo las relaciones obtenidas para los residuos, $e_{ij}=y_{ij}-\widehat{y}_i$, siguen siendo ciertas

$$\sum_{i=1}^{d}\sum_{j=1}^{n_i} e_{ij}=0,\quad \bar{\widehat{y}}=\bar{y},\quad \sum_{i=1}^{d}\sum_{j=1}^{n_i} e_{ij}x_{ij}=0,\quad \sum_{i=1}^{d}\sum_{j=1}^{n_i}(\widehat{y}-\bar{y})e_{ij}=0$$

EJERCICIO: Desarrollar la estimación mínimo-cuadrática del modelo y verificar las propiedades de los residuos.

2.2.2. Interpretación Geométrica del Método de Mínimos Cuadrados

El método de mínimos cuadrados tratado anteriormente admite una fácil y simple interpretación geométrica. Para expresar vectorialmente el conjunto de observaciones y predicciones definamos los vectores siguientes

$$y=(y_1,...,y_N)^t,\qquad \mathbf{1}=(1,...,1)^t,\qquad x=(x_1,...,x_N)^t$$

$$\varepsilon=(\epsilon_1,...,\epsilon_N)^t,\qquad e=(e_1,...,e_N)^t,\qquad \widehat{y}=(\widehat{y}_1,...,\widehat{y}_N)^t$$

Con esta notación el modelo queda en la forma $y=\beta_0\mathbf{1}+\beta_1 x+\varepsilon$

Estimar los parámetros por mínimos cuadrados ordinarios requiere encontrar constantes $\widehat{\beta}_0$ y $\widehat{\beta}_1$ tales que el módulo del vector ε sea mínimo. Por lo tanto se trata de determinar un vector $\widehat{y}$, en el plano definido por los vectores $\mathbf{1}$ y x en un espacio$N$-dimensional, tal que el módulo del vector sea ε sea mínimo.

La solución es tomar la proyección ortogonal del vector y sobre este plano, ya que cualquier otro caso conduciría a un vector residual de módulo mayor. Con ello el vector ε será perpendicular a todos los vectores de dicho plano, bastando para ello con que lo sean con los vectores que lo generan, o sea, los vectores $\mathbf{1}$ y x. Así tendrá lo que se traduce en

$$\varepsilon'\mathbf{1}=0\Leftrightarrow\sum_{i=1}^{N}\epsilon_i=0$$

$$\varepsilon' x=0\Leftrightarrow\sum_{i=1}^{N}\epsilon_i x_i=0$$

lo cual conduce al sistema de ecuaciones normales obtenido anteriormente.

Notemos asimismo que el triángulo resultante de la proyección indica que $\parallel y\parallel^2=\parallel e\parallel^2+\parallel\widehat{y}\parallel^2$ o, lo que es lo mismo,

$$\sum_{i=1}^{N} y_i^2=\sum_{i=1}^{N} e_i^2+\sum_{i=1}^{N}\widehat{y}_i^2$$

2.2.3. Propiedades de los Estimadores Mínimo-Cuadráticos de la Recta de Regresión

Los estimadores son lineales

A partir de la expresión de $\widehat{\beta}_1$ podemos escribir:

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{Ns_x^2} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})y_i}{Ns_x^2} = \sum_{i=1}^{N} w_i y_i, \quad \text{donde} \quad w_i = \frac{x_i-\bar{x}}{Ns_x^2}$$

Observamos por tanto que $\widehat{\beta}_1$ es un **estimador residual** en las variables y_i
Usando la expresión del modelo, tenemos

$$\sum_{i=1}^{N} w_i y_i = \sum_{i=1}^{N}(\beta_0+\beta_1 x_i+\epsilon_i) = \beta_1 + \sum_{i=1}^{N} w_i\epsilon_i \qquad \sum_{i=1}^{N} w_i = 0,\ \sum_{i_1}^{N} w_i x_i = 1$$

Así pues, $\widehat{\beta}_1$ también puede expresarse como **combinación lineal** de las variables de perturbación ϵ_i.
Por otro lado

$$\widehat{\beta}_0 = \bar{y}-\widehat{\beta}_1\bar{x} = \frac{1}{N}\sum_{i=1}^{N} y_i - \bar{x}\sum_{i=1}^{N} w_i y_i = \sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right] y_i = \sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right](\beta_0+\beta_1 x_i+\epsilon_i) = \beta_0+\sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right]\epsilon_i$$

donde se ha usado ahora que $\sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right] = 1$ y $\sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right] x_i = 0$

Por tanto, $\widehat{\beta}_0$ también es combinación lineal tanto de las variables y_i como de las perturbaciones ϵ_i.

Los estimadores son insesgados

De las relaciones anteriores se deduce inmediatamente la insesgadez de ambos estimadores, esto es $E[\widehat{\beta}_0] = \beta_0$ y $E[\widehat{\beta}_1] = \beta_1$

Varianzas y covarianza de los estimadores

Como $\widehat{\beta}_1$ es un estimador insesgado para β_1 se tiene

$$Var[\widehat{\beta}_1] = E[(\widehat{\beta}_1-\beta_1)^2] = E\left[\left(\sum_{i=1}^{N} w_i\epsilon_i\right)^2\right] = E\left[\sum_{i=1}^{N} w_i^2\epsilon_i^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{N} w_i w_j \epsilon_i\epsilon_j\right] = \sum_{i=1}^{N} w_i^2 E[\epsilon_i^2] + \sum_{\substack{i,j=1\\ i\neq j}}^{N} w_i w_j E[\epsilon_i\epsilon_j] =$$

$$= \sigma^2\sum_{i=1}^{N} w_i^2 = \sigma^2\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{N^2(s_x^2)^2} = \frac{\sigma^2}{Ns_x^2}$$

- El error de estimación es inversamente proporcional a la dispersión de los valores de la variable X
- El error de estimación es directamente proporcional a σ^2 o variabilidad intrínseca de la variable respuesta para cada de X fijo

Razonando de forma análoga a la anterior se tiene

$$Var[\widehat{\beta}_0] = E[(\widehat{\beta}_0-\beta_0)^2] = E\left[\left(\sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right]\epsilon_i\right)^2\right] = \cdots = \sigma^2\left[\frac{1}{N}+\frac{\bar{x}^2}{Ns_x^2}\right]$$

Esta expresión muestra la varianza de $\widehat{\beta}_0$ como suma de dos términos: el primero es el error de estimación de la media de la variable Y, mientras que el segundo tiene en cuenta que el error de la estimación de la pendiente de la recta se transmite a la ordenada en el origen (recuérdese cómo se ha obtenido este estimador) en función de lo alejado que se encuentre $\bar{x}$ del origen.

EJERCICIO: Realizar el desarrollo del cálculo de $Var[\widehat{\beta}_0]$.

En cuanto a la covarianza tenemos

$$Cov[\widehat{\beta}_0,\widehat{\beta}_1] = E[(\widehat{\beta}_0-\widehat{\beta}_0)(\widehat{\beta}_1-\widehat{\beta}_1)] = E\left[\left(\sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right]\epsilon_i\right)\left(\sum_{j=1}^{N}w_j\epsilon_j\right)\right]$$
$$= E\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right]w_j\epsilon_i\epsilon_j\right] = \sum_{i=1}^{N}\sum_{j=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right]w_jE[\epsilon_i\epsilon_j]$$
$$= \sigma^2\sum_{i=1}^{N}\left[\frac{1}{N}-\bar{x}w_i\right]w_i = \sigma^2\sum_{i=1}^{N}\left[\frac{x_i-\bar{x}}{N^2s_x^2}-\bar{x}\frac{(x_i-\bar{x})^2}{N^2(s_x^2)^2}\right] = -\sigma^2\frac{\bar{x}}{Ns_x^2}$$

Con ello se observa que si $\bar{x}$ es mayor que cero entonces errores por exceso en la estimación de la pendiente produce errores por defecto en la estimación de la ordenada en el origen y recíprocamente ocurre otro tanto.

Eficiencia de los Estimadores

Antes hemos comprobado cómo los estimadores mínimo cuadráticos de β_0 y β_1 son lineales en las observaciones de la variable dependiente (y también en las realizaciones de la variable de perturbación), y además son insesgados.

Se puede demostrar que esos estimadores cumplen la propiedad de que entre todos los posibles estimadores lineales e insesgados que se puedan obtener para los parámetros (no tienen por qué existir solo ellos), son los que tienen menor varianza.

Así pues son los estimadores de mínima varianza de entre todos los posibles estimadores lineales e insesgados de los parámetros (**BLUE**). Este resultado es una versión *light* del Teorema de Gauss-Markov.

NOTA: La demostración se hará como complemento en la clase de grupo pequeño.

2.2.4. Estimación de σ^2. Varianza Residual

Nos planteamos a continuación la estimación de la varianza de las perturbaciones, σ^2. Con esa cantidad dispondremos de información sobre cuanta variabilidad del modelo deja de explicarse. Para ello partimos de los residuos estimados de la regresión $e_i = y_i - \widehat{y}_i$, ya que se pueden interpretar como realizaciones de las variables de perturbación.

Recordando que $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, se tiene $e_i = (\beta_0-\widehat{\beta}_0)+(\beta_1-\widehat{\beta}_1)x_i+\epsilon_i,\ i=1,...,N$, con lo cual $E[e_i]=0,\ i=1,...,N$.

Consideremos ahora la esperanza de los cuadrados de los residuos

$$E[e_i^2] = E\left[(\beta_0-\widehat{\beta}_0)^2+(\beta_1-\widehat{\beta}_1)^2x_i^2+\epsilon_i^2+2(\beta_0-\widehat{\beta}_0)(\beta_1-\widehat{\beta}_1)x_i+2(\beta_0-\widehat{\beta}_0)\epsilon_i+2(\beta_1-\widehat{\beta}_1)x_i\epsilon_i\right]$$
$$= Var[\widehat{\beta}_0]+Var[\widehat{\beta}_1]x_i^2+Var[\epsilon_i]+2Cov[\widehat{\beta}_0,\widehat{\beta}_1]x_i-2E[(\beta_0-\widehat{\beta}_0)\epsilon_i]-2E[(\beta_1-\widehat{\beta}_1)x_i\epsilon_i]$$

y como

$$E[(\widehat{\beta}_0-\beta_0)\epsilon_i] = \frac{\sigma^2}{N}-\frac{\bar{x}(x_i-\bar{x}\sigma^2}{Ns_x^2} \quad \text{y} \quad E[(\widehat{\beta}_1-\beta_1)x_i\epsilon_i] = \frac{\sigma^2(x_i-\bar{x})x_i}{Ns_x^2}$$

se concluye

$$E[e_i^2] = \sigma^2\left[1-\frac{1}{N}-\frac{(x_i-\bar{x})^2}{Ns_x^2}\right] = (N-2)\sigma^2,$$

de donde se deduce que un estimador insesgado de la varianza(la varianza residual) es

$$\widehat{\sigma}^2 = \sum_{i=1}^{N}\frac{e_i^2}{(N-2)}$$

EJERCICIO: Completar los cálculos anteriores y realizar la estimación de σ^2 para el modelo sin término constante.

2.3. Estimación del Modelo por Máxima Verosimilitud

Como ta se ha comentado anteriormente, la hipótesis de normalidad sobre los términos de error conlleva el hecho de que las variables y_i sean normales e independientes, por lo que es inmediato construir la función de

verosimilitud asociada a la muestra $\mathbf{y} = \{y_1, ..., y_N\}$.

$$\mathbb{L}_{\mathbf{y}}(\beta_0, \beta_1, \sigma^2) = (2\pi\sigma^2)^{\frac{N}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[y_i - \beta_0 - \beta_1 x_i]^2\right)$$

Nuestra idea es obtener los estimadores máximo-verosímiles para los parámetros β_0, β_1 y σ^2. Para ello hay que realizar la maximización de la anterior función y encontrar $\tilde{\beta}_0, \tilde{\beta}_1$ y $\tilde{\sigma}^2$ tales que

$$\mathbb{L}_{\mathbf{y}}(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2) = \underset{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2}{\text{Sup}} \mathbb{L}_{\mathbf{y}}(\beta_0, \beta_1, \sigma^2)$$

Para ello, en principio, habrá que derivar la función de verosimilitud y encontrar sus puntos críticos. Asimismo, como trabajar con dicha función es algo complicado, lo que suele hacerse es considerar su logaritmo ya que, al ser el logaritmo creciente, conserva los puntos críticos. En este caso

$$log(\mathbb{L}_{\mathbf{y}}(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2)) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}[y_i - \beta_0 - \beta_1 x_i]^2$$

Derivando respecto de los parámetros, se tiene

$$\frac{\partial log\mathbb{L}_{\mathbf{y}}(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2)}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{N}[y_i - \beta_0 - \beta_1 x_i] \qquad \frac{\partial log\mathbb{L}_{\mathbf{y}}(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2)}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{N}[y_i - \beta_0 - \beta_1 x_i]x_i$$

$$\frac{\partial log\mathbb{L}_{\mathbf{y}}(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}[y_i - \beta_0 - \beta_1 x_i]^2$$

Igualando las dos primeras parciales a cero se tiene el siguiente sistema de ecuaciones

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

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

que no es otro que el sistema de ecuaciones normales que se obtuvo tras la estimación por mínimos cuadrados ordinarios. Con ello los estimadores máximo verosímiles de β_0 y β_1 son los mismos que los mínimo cuadráticos.

A continuación igualamos la tercera parcial a cero y sustituimos los parámetros β_0 y β_1 por los estimadores anteriores, concluyendo que el estimador máximo verosímil para σ^2 es

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{N}$$

que ya no coincide con el estimador de mínimos cuadrados, si bien están relacionados:

$$\tilde{\sigma}^2 = \frac{N-2}{N}\widehat{\sigma}^2$$

Además este estimador no es insesgado ya que

$$E[\tilde{\sigma}^2] = \frac{N-2}{N}\sigma^2$$

2.4. Distribución de los Estimadores

El hecho de que los estimadores β_0 y β_1 se puedan expresar en términos de las perturbaciones de ϵ_i y que éstas sean, por hipótesis, variables aleatorias independientes e idénticamente distribuidas según una normal de media cero y varianza σ^2, permite obtener de forma rápida las distribuciones de dichos estimadores. En cuanto a la distribución de la varianza residual no tendremos una expresión para ella si bien, aunque lo demostraremos más adelante, sí es conocida la distribución de una cierta función suya.

2.4.1. Distribución de $\widehat{\beta_1}$

Recordemos que $\widehat{\beta_1}$ se puede expresar como

$$\widehat{\beta_1} = \beta_1 + \sum_{i=1}^{N} w_i \epsilon_i,$$

con lo cual se distribuirá según una normal de media β_1 y de varianza

$$\sum_{i=1}^{N} w_i^2 \sigma^2 = \frac{\sigma^2}{N s_x^2}$$

puesto que es una combinación lineal de variables normales, independientes e idénticamente distribuidas.

2.4.2. Distribución de $\widehat{\beta_0}$

En este caso, como

$$\widehat{\beta_0} = \beta_0 + \sum_{i=1}^{N} \left[\frac{1}{N} - \bar{x} w_i \right] \epsilon_i$$

su distribución es normal de media β_0 y varianza

$$\sigma^2 \sum_{i=1}^{N} \left[\frac{1}{N} - \bar{x} w_i \right]^2 = \sigma^2 \left[\frac{1}{N} + \frac{\bar{x}^2}{N S_x^2} \right]$$

2.4.3. Distribución de $\widehat{\sigma^2}$

La distribución del estimador de la varianza de los términos de error no es inmediata puesto que se basa en la distribución de formas cuadráticas normales. Por esta razón omitimos su demostración aquí, si bien posteriormente trataremos esta cuestrión más detalladamente. Concretamente se tiene que

$$\frac{\sum_{i=1}^{N} e_i^2}{\sigma^2} = \frac{(N-2)\widehat{\sigma}^2}{\sigma^2}$$

se distribuye según una distribución χ^2 centrada con $N-2$ grados de libertad.

Los $N-2$ grados de libertad de los residuos proceden del número de datos menos el número de parámetros que han hecho falta estimar para calcular las medias en cada punto, pudiéndose entender mejor si consideramos que los N residuos no son independientes ya que los residuos poseen dos restricciones dadas por

$$\sum_{i=1}^{N} e_i = 0 \quad , \quad \sum_{i=1}^{N} e_i x_i = 0$$

con lo cual hay sólo $N-2$ residuos independientes, que corresponden a los $N-2$ grados de libertad.

Además puede demostrarse que tanto $\widehat{\beta_0}$ como $\widehat{\beta_1}$ son independientes de $\sum_{i=1}^{N} e_i^2$ (si bien ello conlleva estudiar independencia de formas cuadráticas y lineales y lo dejaremos par el apartado de regresión múltiple).

2.5. Descomposición de la variabilidad. Coeficiente de Determinación

2.5.1. Descomposición de la Variabilidad

Una vez estimados los parámetros del modelo nos planteamos la bondad de tal estimación.

Ello se traduce en averiguar hasta qué punto el modelo es capaz de reproducir el comportamiento de la variable dependiente, que es el objeto del estudio; es decir, qué parte de la variación de la variable dependiente viene explicada por la recta estimada.

Para ello hemos de proceder a descomponer la variabilidad de la variable explicada e intentar darle un significado a dicha descomposición(cuestión que es común a todos los modelos lineales). Lo haremos para el modelo con término constante.

Sabemos que $y_i = \widehat{y}_i + e_i,\ i = 1, ..., N$, de donde $y_i - \bar{y} = \widehat{y}_i - \bar{y} + e_i,\ i = 1, ..., N$. Elevando al cuadrado

$$(y_i - \bar{y})^2 = (\widehat{y}_i - \bar{y})^2 + e_i^2 + 2(\widehat{y}_i - \bar{y})e_i; \quad i = 1, ..., N$$

y sumando en ambos miembros

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{N} e_i^2 \qquad \left(\text{ya que } \sum_{i=1}^{N}(\widehat{y}_i - \bar{y})e_i = 0\right)$$

Esta expresión descompone la variabilidad de la variable dependiente (VT) en dos términos: el segundo de ellos refleja la parte de variabilidad que no es explicada por la regresión puesto que es la variabilidad de los residuos, que llamaremos variabilidad no explicada (VNE), mientras que el primero de ellos refleja toda la variabilidad de la variable dependiente que ha sido explicada por la regresión (observemos cómo ese primer término no es más que la variabilidad de la variable predicha puesto que $\bar{\widehat{y}} = \bar{y}$) y la denominaremos variabilidad explicada (VE).

NOTA: En el caso de modelo sin término constante la descomposición es la dada en la interpretación geométrica del método de mínimos cuadrados, es decir $\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} e_i^2 + \sum_{i=1}^{N} \widehat{y}_i^2$.

Merece especial atención el caso de datos repetidos puesto que en este ambiente puede comprobarse cómo la parte residual se divide a su vez en otros dos términos.

La idea es que el hecho de que para cada valor de x_i, la distribución de Y se sustituya por la media condicionada $\bar{y}_i$, debe traducirse en la introducción de un error adicional. Ese error debe ser el de la representatividad de dicha media condicionada en esa distribución. Por lo tanto es esperable la existencia de dos fuentes de error: una debida a la falta de ajuste de la media condicionada a la recta, y otra debida a la variabilidad de cada distribución condicionada entorno a su media.

Razonando igual que antes

$$y_{ij} = \widehat{y}_i + e_{ij} \quad ; \quad i = 1, ..d \quad , \quad j = 1, ..., n_i$$

de donde

$$y_{ij} - \bar{y} = \widehat{y}_i - \bar{y} + e_{ij} \quad ; \quad i = 1, ..d \quad , \quad j = 1, ..., n_i$$

Elevando al cuadrado y sumando en ambos miembros se tiene

$$\sum_{i=1}^{d}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{d} n_i(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{d}\sum_{j=1}^{n_i} e_{ij}^2 \qquad \left(\sum_{i=1}^{d}\sum_{j=1}^{n_i}(\widehat{y}_i - \bar{y})e_{ij} = 0\right)$$

Esta es la descomposición obtenida anteriormente. A continuación veamos que la variabilidad residual puede descomponerse a su vez. Partamos de los residuos

$$e_{ij} = y_{ij} - \widehat{y}_i = y_{ij} - \bar{y}_i + \bar{y}_i - \widehat{y}_i \quad ; \quad i = 1, ..., d \quad , \quad j = 1, ..., n_i$$

de donde

$$e_{ij}^2 = (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \widehat{y}_i)^2 + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \widehat{y}_i) \quad ; \quad i = 1, ..., d \quad , \quad j = 1, ..., n_i$$

y sumando en ambos miembros

$$\sum_{i=1}^{d}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{d}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{d}\sum_{j=1}^{n_i}(\bar{y}_i - \widehat{y}_i)^2$$

donde se ha usado que

$$\sum_{i=1}^{d}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(\bar{y}_i - \widehat{y}_i) = \sum_{i=1}^{d}(\bar{y}_i - \widehat{y}_i)\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i) = 0$$

Con ello tenemos

$$\text{VNE} = \sum_{i=1}^{d}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{d} n_i(\bar{y}_i - \widehat{y}_i)^2$$

Así pues la variabilidad residual se descompone en dos términos. El segundo marca las diferencias al cuadrado entre los valores predichos, para cada valor x_i de la variable explicativa, y el valor que se prediciría sin la necesidad de la recta(media condicionada $\bar{y}_i$) y por lo tanto determina la falta de ajuste de la nube de puntos a la recta estimada. El primero indica las diferencias, dentro de cada grupo, de los valores observados y su media condicionada y, por lo tanto, es un error intrínseco a los datos que viene dado por representar cada distribución condicionada $y|x_i$ por medio de su media condicionada.

En resumen, la variabilidad residual es el compendio del error intrínseco a los datos y el determinado por la falta de linealidad, siendo esta descomposición la base del contraste de linealidad que se verá posteriormente.

2.5.2. Coeficiente de Determinación

La evaluación global de la recta estimada se puede hacer mediante la varianza residual, que es un índice de la precisión del modelo. Sin embargo esta medida no es útil para tal fin puesto que depende de las unidades de medida en las que venga expresada la variable dependiente. Una medida más adecuada del ajuste sería la proporción de variabilidad explicada frente a la variabilidad total, con lo cual tendremos un valor adimensional que nos dirá cuánta parte de la variabilidad total es explicada por el modelo.

Así se define el coeficiente de determinación como

$$R^2 = \frac{\text{VE}}{\text{VT}} = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{\text{VNE}}{\text{VT}} = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Es inmediato comprobar que dicho coeficiente toma valores entre 0 y 1. Además, cuanto más próximo esté a uno indicará que mayor es la parte de variabilidad de la variable dependiente que queda explicada por la regresión, siendo éste un indicio de un buen comportamiento de la estimación. Por el contrario, cuanto más próximo esté a cero indicará que el modelo estimado no es óptimo para explicar el comportamiento de la variable explicada.

Asimismo se puede demostrar que, en el caso de la regresión lineal simple univariante, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación lineal entre las variables explicada y explicativa.

En efecto, de $\widehat{y}_i - \bar{y} = \widehat{\beta}_1(x_i - \bar{x}),\ i = 1, ..., N$, se tiene

$$(\widehat{y}_i - \bar{y})^2 = \widehat{\beta}_1^2(x_i - \bar{x})^2 \quad ; \quad i = 1, ..., N$$

de donde

$$\sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2 = \widehat{\beta}_1^2 \sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{N s_{xy}^2}{s_x^2} \qquad \left(\text{VE} = \widehat{\beta}_1^2 N s_x^2\right)$$

Con ello

$$R^2 = \frac{\text{VE}}{\text{VT}} = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{\frac{N s_{xy}^2}{s_x^2}}{N s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} = r_{xy}^2$$

A partir de esta identidad es inmediato deducir:

- $0 \leq R^2 \leq 1$
- $R^2 = 0 \Leftrightarrow S_{xy} = 0 \Leftrightarrow r = 0$. En tal caso se tiene que no existe dependencia lineal entre ambas variables, por lo que el conocimiento de X no aporta ninguna información a la estimación de posibles valores de Y. En tal caso la única estimación posible sobre el comportamiento de Y sería su media, quedando la recta de regresión en la forma $\widehat{y} = \bar{y}$
- $R^2 = 1 \Leftrightarrow r_{xy} \pm 1$. En este caso la regresión entre X e Y es exacta y hay una dependencia lineal completa entre ambas variables

EJERCICIO: ¿Cuál es la expresión de R^2 en el modelo sin término constante?

2.5.3. Distribución de las Variabilidades Explicada y no Explicada

Puesto que $\text{VNE} = \sum_{i=1}^{N} e_i^2$, entonces $\frac{\text{VNE}}{\sigma^2} \rightsquigarrow \chi^2_{N-2}$

Por otro lado se verifica que $\text{VE} = \sum_{i=1}^{N}(\widehat{y}_i \bar{y})^2 = \widehat{\beta}_1^2 N S_x^2$

Como $\widehat{\beta}_1 \rightsquigarrow N_1\left[\beta_1; \frac{\sigma^2}{Ns_x^2}\right]$, entonces $\sqrt{NS_x^2}\widehat{\beta}_1 \rightsquigarrow N_1\left[\sqrt{NS_x^2}\beta_1; \sigma^2\right]$, por lo que

$$\frac{\text{VE}}{\sigma^2} \rightsquigarrow \chi_1^2(\delta) \quad \text{siendo} \quad \delta = \frac{Ns_x^2\beta_1^2}{\sigma^2}$$

Además, puesto que $\widehat{\beta}_1$ y $\widehat{\sigma}^2$ son independientes se deduce que ambas variables son independientes y que

$$\frac{\text{VE}}{\frac{\text{VNE}}{N-2}} \rightsquigarrow F_{1,N-2}(\delta)$$

2.6. Predicción

Una de las utilidades de un modelo de regresión es predecir el comportamiento de la variable dependiente conocidos los valores de la variable explicativa. Dicho conocimiento puede referirse bien a la estimación de la media de la distribución de la variable explicada para cada nuevo valor de la variable independiente o bien predecir futuros valores de dicha variable.

Dicho en otras palabras, si x_p es un nuevo valor de la variable explicativa(no ha contribuido en el proceso de estimación), se pretende o bien obtener el valor predicho de la variable dependiente conocido x_p ($\widehat{y}_p$) o bien estimar el valor medio de la variable dependiente, conocido x_p ($\widehat{E[y_p]}$).

Ambas cuestiones se resuelven de idéntica manera: sustituyendo en la recta de regresión el valor dado de la variable explicativa, puesto que la predicción mediante $\widehat{y}_p$ es insesgada. En efecto, el valor exacto de la variable dependiente sería

$$y_p = \beta_0 + \beta_1 x_p + \epsilon_p$$

donde ϵ_p es una perturbación que verifica las mismas hipótesis que las del modelo. Así el valor medio esperado para y_p sería

$$E[y_p] = \beta_0 + \beta_1 x_p$$

que habría que estimar pues viene en función de los parámetros. Dicha estimación sería

$$\widehat{E[y_p]} = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$$

que coincide, obviamente, con el valor que la recta predice para la variable dependiente.

No obstante la precisión(medida mediante el error cuadrático medio) no es la misma en los dos casos, como vemos a continuación.

Precisión en el Caso de la Estimación de la Media

En este caso el error cuadrático medio viene dado por la varianza del estimador $\widehat{E[y_p]}$. Como

$$\widehat{E[y_p]} - E[y_p] = (\widehat{\beta}_0 - \beta_0) + (\widehat{\beta}_1 - \beta_1)x_p$$

se tiene

$$\begin{aligned} Var[\widehat{E[y_p]}] &= E[(\widehat{E[y_p]} - E[y_p])^2] = Var[\widehat{\beta}_0] + x_p^2 Var[\widehat{\beta}_1] + 2x_p Cov[\widehat{\beta}_0, \widehat{\beta}_1] \\ &= \sigma^2\left\{\left[\frac{1}{N} + \frac{\bar{x}^2}{Ns_x^2}\right] + \frac{x_p^2}{Ns_x^2} - 2\frac{x_p\bar{x}}{Ns_x^2}\right\} = \sigma^2\left[\frac{1}{N} + \frac{(\bar{x} - x_p)^2}{Ns_x^2}\right] \end{aligned}$$

expresión que depende de la distancia del punto x_p a la media de la variable explicativa.

Precisión en el Caso de la Predicción de un Nuevo Valor

Supongamos ahora que deseamos predecir el valor de la variable dependiente (y_p) para un valor dado (x_p) de la variable independiente. Evidentemente dicho valor yp no ha contribuido en la estimación de los parámetros del modelo. Dicha predicción se efectúa(al igual que en el caso de la estimación de la media de la variable respuesta) sustituyendo el valor de x_p en la recta de regresión estimada.

No obstante, el error de la predicción vendrá ahora dado por la desviación cuadrática media de la predicción a la variable respuesta, o sea $E[(y_p - \widehat{y}_p)^2]$. Ahora bien,

$$\begin{aligned} E[(y_p - \widehat{y}_p)^2] &= E\left[\left\{(\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1)x_p + \epsilon_p\right\}^2\right] \\ &= Var[\widehat{\beta}_0] + Var[\widehat{\beta}_1]x_p^2 + Var[\epsilon_p] + 2x_p Cov[\widehat{\beta}_0, \widehat{\beta}_1] - 2Cov[\widehat{\beta}_0, \epsilon_p] - 2x_p Cov[\widehat{\beta}_1, \epsilon_p] \\ &= \sigma^2\left\{\left[\frac{1}{N} + \frac{\bar{x}^2}{Ns_x^2}\right] + \frac{x_p^2}{Ns_x^2} + 1 - \frac{2x_p\bar{x}}{Ns_x^2}\right\} = \sigma^2\left[1 + \frac{1}{N} + \frac{(\bar{x} - x_p)^2}{Ns_x^2}\right] \end{aligned}$$

donde se ha usado que y_p es independiente de $\widehat{\beta}_0$ y $\widehat{\beta}_1$ pues no intervino en su estimación.

Observamos que la precisión de la predicción es la suma de la varianza de la estimación de la media de la variable dependiente y la varianza de la propia variable alrededor de su media teórica.