Capítulo 4

Introducción a los modelos lineales

En lo que sigue, centraremos nuestro interés en relaciones del tipo:

$$E(Y) = f(X_1, \ldots, X_n)$$

donde $X_1, \ldots, X_n$ son valores de variables explicativas y $E(Y)$ es el valor promedio de cierta respuesta.

Cuando consideramos que la transmisión y recepción de información involucra en general un mensaje o señal que es distorsionado por un ruido, es útil pensar que la señal es determinística, y que el ruido es aleatorio. En este caso, hablamos de un *modelo probabilístico o estadístico.* Los modelos que estudiaremos a continuación tienen la forma

$$Y = \text{señal} + \text{ruido},$$

es decir

$$Y = f(X_1, \ldots, X_n) + \varepsilon$$

donde ε es una variable aleatoria.

Consideremos a continuación algunas situaciones en las cuales es razonable usar este tipo de modelos.

Ejemplo 4.1 *Supongamos que semillas genéticamente similares son asignadas aleatoriamente a dos ambientes, uno enriquecido (tratamiento) y uno standard (control). Luego de un período determinado, las plantas son cosechadas y pesadas. Los resultados se muestran en la tabla 4.1.*

¿ Existen diferencias entre el crecimiento de las plantas en el ambiente enriquecido y el ambiente standard?

Control	Tratamiento
4.87	4.93
4.67	4.36
4.40	5.35
4.88	4.02
4.86	4.66
5.14	4.82
3.86	3.68
4.71	4.57
5.33	4.19
5.56	4.22

Tabla 4.1: Peso de plantas cultivadas en dos ambientes enriquecido y standard (Ejemplo 4.1)

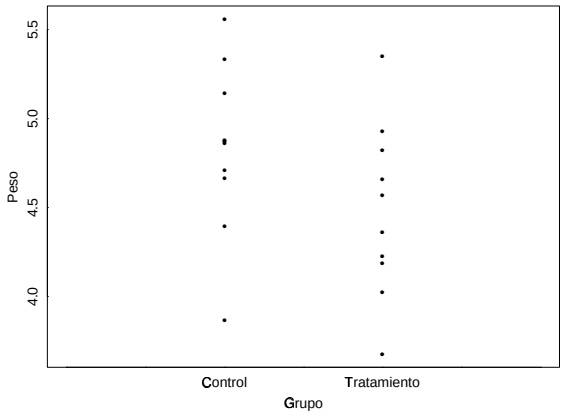

Figura 4.1: Gráfico para el ejemplo de crecimiento de plantas.

Cuando graficamos estos datos, obtenemos el dibujo que se presenta en la figura 4.1.

En base a este dibujo, parece razonable pensar que para cada uno de los grupos (control y tratamiento) existe un valor central alrededor del cual se distribuyen las observaciones, y que dicho valor es distinto en cada caso. En base a lo anterior, planteamos entonces el siguiente modelo:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2; \quad j = 1, \ldots, 10 \tag{4.1}$$

donde

Y_{ij}	:	Peso de la planta j del grupo i.
μ_i	:	Peso promedio para el grupo i.
ε_{ij}	:	Componente de ruido (se considera aleatoria).

μ_1 y μ_2 serán los *parámetros del modelo*, es decir, aquellas cantidades cuyo valor determina la parte de señal del modelo en estudio, y las cuales son claramente desconocidas. Por esta razón, necesitamos usar alguna función de los datos que nos permita tener una idea de los valores de estos parámetros. A este procedimiento lo denominamos *estimación* de los parámetros del modelo, y a las funciones de los datos que emplearemos las llamaremos *estimadores*.

Es de interés probar la hipótesis de que ambos ambientes no generan diferencia. Si es éste el caso, debemos comparar el modelo anterior con el modelo que contempla una única media para todos los datos, es decir:

$$y_{ij} = \mu + \varepsilon_{ij}, \quad i = 1, 2; \quad j = 1, \ldots, 10 \tag{4.2}$$

En resumen, se nos plantean dos problemas fundamentales:

- Cómo estimar los parámetros del modelo.
- Cómo comparar modelos.

Ejemplo 4.2 *En la tabla 4.2 están los pesos en gramos al nacer y la edad gestacional en semanas de un grupo de niños y niñas nacidos en cierto hospital. Es de interés determinar si la edad gestacional puede ayudar a predecir el peso al nacer, y si existen diferencias entre niños y niñas.*

Graficando estos datos obtenemos el resultado que se muestra en la figura 4.2. Observamos que el peso al nacer parece aumentar en forma lineal con la edad gestacional. Ahora bien, ¿será igual este aumento para las niñas que para los niños?

Varones		Hembras	
Edad Gestacional (semanas)	Peso al nacer (Kgs.)	Edad Gestacional (semanas)	Peso al nacer (Kgs.)
40	2.968	40	3.317
38	2.795	36	2.724
40	3.165	40	2.935
35	2.925	38	2.754
36	2.625	42	3.210
37	2.847	39	2.817
41	3.292	40	3.126
40	3.473	37	2.539
37	2.628	36	2.412
38	3.176	38	2.991
40	3.421	39	2.875
38	2.975	40	3.231

Tabla 4.2: Edad gestacional y peso al nacer para un grupo de bebés (Ejemplo 4.2)

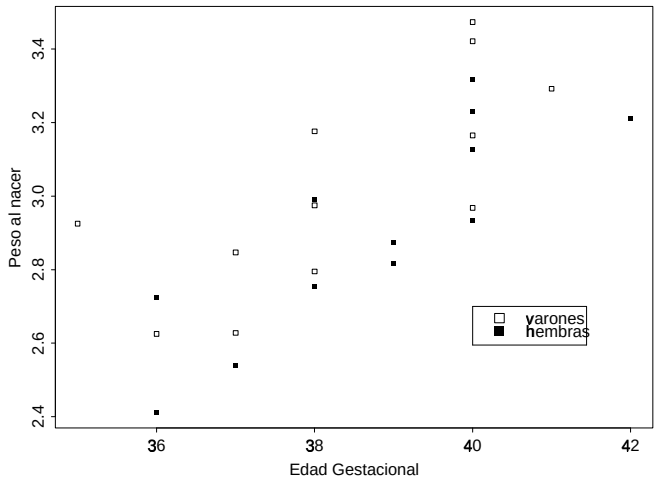

Figura 4.2: Peso al nacer y edad gestacional para niñas y niños.

Un modelo bastante general para la situación antes descrita es el siguiente:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

con $i = 1$ para los varones, $i = 2$ para las hembras y j variando entre 1 y 12. En este modelo, cada término representa:

Y_{ij}	:	Peso al nacer del j-ésimo bebe de sexo i.
α_i	:	ordenada en el origen para bebés de sexo i.
β_i	:	pendiente para bebés de sexo i.
x_{ij}	:	edad gestacional del j-ésimo bebe de sexo i.
ε_{ij}	:	Componente de ruido (se considera aleatoria).

Sin embargo, otros modelos más sencillos son posibles. Por ejemplo, podemos pensar que la ordenada en el origen es igual para varones y hembras, lo cual equivale al modelo:

$$Y_{ij} = \alpha + \beta_i x_{ij} + \varepsilon_{ij}.$$

Si decidimos que no hay diferencias en la pendiente de las rectas correspondientes a varones y hembras, el modelo correspondiente será:

$$Y_{ij} = \alpha_i + \beta x_{ij} + \varepsilon_{ij}.$$

Para representar la creencia de que el comportamiento de niños y niñas es el mismo, recurrimos al modelo:

$$Y_{ij} = \alpha + \beta x_{ij} + \varepsilon_{ij}.$$

Pueden incluirse otras simplificaciones, las cuales a su vez equivalen a modelos diferentes que pueden ajustarse a la situación estudiada. Nuevamente, se nos presentan los problemas de estimación de los parámetros del modelo y de comparación de los modelos posibles para decidir cuál de ellos se ajusta mejor a los datos obtenidos.

Ejemplo 4.3 *La resistencia a la tensión de cierto papel está relacionada con la cantidad de cierta madera que está presente en la pulpa. Se toman 10 muestras y se obtienen los datos que se muestran en la tabla 4.3. ?'Cómo influye el porcentaje de madera sobre la resistencia del papel?*

Si graficamos la resistencia obtenida versus el porcentaje de madera, obtenemos el resultado que se muestra en la figura 4.3, el cual da buenas razones para pensar que la dependencia es lineal. Podemos plantear, entonces, un modelo de la forma:

Resistencia	160	171	175	182	184	181	188	193	195	200
% Madera	10	15	15	20	20	20	25	25	28	30

Tabla 4.3: Resistencia a la tensión y porcentaje de madera para 10 muestras de papel (Ejemplo 4.3)

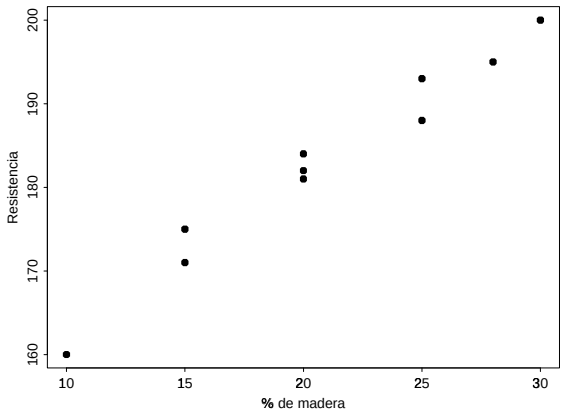

Figura 4.3: Resistencia a la tensión vs. Porcentaje de madera.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Para simplificar el modelo, podríamos plantear la hipótesis de que el porcentaje de madera no influye sobre la resistencia a la tensión en el papel producido. Esta hipótesis corresponde al modelo:

$$Y_i = \beta_0 + \varepsilon_i$$

Nuevamente, tenemos el problema de cómo estimar los parámetros, y de cómo comparar los modelos anteriores.

4.1 Modelo lineal

Veamos que estos modelos pueden englobarse en una clase más amplia. Para ello, los escribiremos en forma matricial.

El modelo 4.1 en el ejemplo 4.1 puede ser escrito en forma matricial como:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

donde

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1,10} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2,10} \end{pmatrix} = \begin{pmatrix} 4.87 \\ 4.67 \\ \vdots \\ 5.56 \\ 4.93 \\ 4.36 \\ \vdots \\ 4.22 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1,10} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2,10} \end{pmatrix}$$

Es decir, $\mathbf{Y}$ es un vector que contiene los pesos de las plantas para los dos grupos, $\boldsymbol{\beta}$ es el vector de parámetros, X es una matriz de indicadores y $\boldsymbol{\varepsilon}$ es el vector de errores.

Para el ejemplo 4.2, el primer modelo planteado puede escribirse nuevamente en forma matricial como:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

definiendo cada uno de los elementos del modelo de la siguiente manera:

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1,12} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2,12} \end{pmatrix} = \begin{pmatrix} 2.968 \\ 2.795 \\ \vdots \\ 2.975 \\ 3.317 \\ 2.724 \\ \vdots \\ 3.231 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 40 & 0 \\ 1 & 0 & 38 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 38 & 0 \\ 0 & 1 & 0 & 40 \\ 0 & 1 & 0 & 36 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 40 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1,12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2,12} \end{pmatrix}$$

En este caso, $\mathbf{Y}$ contiene los pesos al nacer para varones y hembras, la matriz X contiene indicadores y los valores de la edad gestacional para cada bebé, $\boldsymbol{\beta}$ es el vector de parámetros y $\boldsymbol{\varepsilon}$ es nuevamente el vector de errores.

Podemos englobar estos modelos dentro de una clase general de modelos de la forma:

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.3}$$

donde

- $\mathbf{Y}$: Vector $n \times 1$ de respuestas (aleatorias)
- X : Matriz $n \times p$ que contiene ceros, unos y/o valores de variables independientes
- $\boldsymbol{\beta}$: Vector $p \times 1$ de parámetros
- $\boldsymbol{\varepsilon}$: vector $n \times 1$ de errores aleatorios.

Un modelo de esta forma se denomina *Modelo Lineal*, ya que la componente de señal del modelo ($X\boldsymbol{\beta}$) es una combinación lineal de los parámetros, y la componente de ruido ($\boldsymbol{\varepsilon}$) es aditiva.

Es importante enfatizar que la linealidad del modelo es *en los parámetros* y no en las variables independientes. Consideremos, por ejemplo, los pares de datos $(x_1, y_1), \ldots, (x_n, y_n)$ y supongamos que se cree que la relación entre y y x puede ser representada por medio de una parábola, es decir, por un modelo de la forma:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Este modelo *sí* es un modelo lineal. En efecto, al escribirlo en forma matricial se obtiene $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, donde

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Lo anterior vale para modelos polinomiales, o para modelos que involucren cualquier transformación de las variables independientes, siempre y cuando la variable dependiente pueda expresarse como función lineal de los parámetros del modelo.

Se supone en general que los errores ε_i tienen media 0, varianza común σ^2 y son independientes entre sí. Es decir,

$$E(\boldsymbol{\varepsilon}) = 0$$

$$Var(\boldsymbol{\varepsilon}) = \sigma^2 I$$

Más adelante fijaremos una distribución de probabilidad para los errores, la cual será necesaria para determinar distribuciones de referencia con el fin de probar hipótesis sobre el modelo.

4.2 Estimación de los parámetros del Modelo Lineal

Consideremos el modelo lineal definido en (4.3)

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Si suponemos que $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, entonces

$$E(\mathbf{Y}) = X\boldsymbol{\beta} = \beta_1 X_1 + ... + \beta_p X_p,$$

siendo X_i la i-ésima columna de X.

Entonces, $E(\mathbf{Y}) \in gen\{X\}$, donde $gen\{X\}$ es el subespacio de R^n generado por las columnas de X. Esto nos lleva a considerar la posibilidad de aproximar $\mathbf{Y}$ usando un elemento de $gen\{X\}$, digamos $\hat{\mathbf{Y}}$. Para ello, parece razonable tomar el elemento de $gen\{X\}$ que está más cerca de $\mathbf{Y}$, es decir, escogeremos $\hat{\mathbf{Y}}$ tal que

$$||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \min_{\mathbf{Z} \in gen\{X\}} ||\mathbf{Y} - \mathbf{Z}||^2$$

A este procedimiento se le conoce como *Método de Mínimos Cuadrados*, pues se busca aquel elemento de $gen\{X\}$ que minimiza la suma de cuadrados de los errores de estimación.

Ésto equivale a estimar el vector de parámetros $\boldsymbol{\beta}$ como el vector $\hat{\boldsymbol{\beta}}$ tal que

$$||\mathbf{Y} - X\hat{\boldsymbol{\beta}}||^2 = \min_{\boldsymbol{\beta} \in R^p} ||\mathbf{Y} - X\boldsymbol{\beta}||^2$$

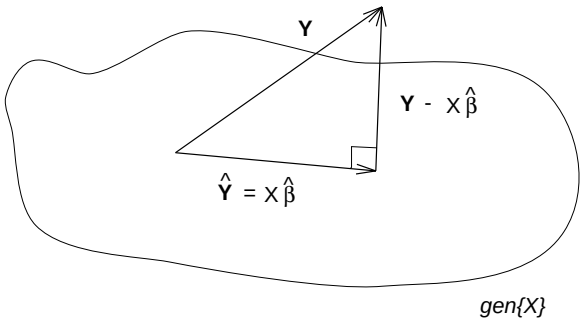

Figura 4.4: Geometría del método de mínimos cuadrados.

Esta igualdad se cumple si y sólo si $\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}}$ es la proyección ortogonal de $\mathbf{Y}$ sobre $gen\{X\}$. En particular, ésto significa que $\mathbf{Y} - X\hat{\boldsymbol{\beta}} \in gen\{X\}^{\perp}$ (ver figura 4.4). Por lo tanto:

$$\begin{aligned} X'(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) &= 0 \\ X'\mathbf{Y} - X'X\hat{\boldsymbol{\beta}} &= 0 \end{aligned}$$

y de esta última expresión se obtiene el sistema de ecuaciones que suele llamarse *Ecuaciones Normales*

$$(X'X)\hat{\boldsymbol{\beta}} = X'\mathbf{Y} \tag{4.4}$$

Es importante destacar que el método de mínimos cuadrados es un método de estimación puramente geométrico, y por tanto es independiente de la estructura probabilística de los errores.

Si $X_{n\times p}$ es de rango máximo (es decir, su rango -número de columnas y/o filas linealmente independientes- es igual a p, el número de columnas), $X'X$ es una matriz invertible. Esto garantiza que $\hat{\boldsymbol{\beta}}$ es único y puede calcularse como:

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y}$$

Si X no es de rango máximo, $X'X$ no es invertible y las ecuaciones normales tienen infinitas soluciones (si bien la proyección $\hat{\mathbf{Y}}$ es única, existen infinitos $\hat{\boldsymbol{\beta}}$ que la generan)

Apliquemos el método de mínimos cuadrados al ejemplo 4.1 (Crecimiento de Plantas) y estimemos los parámetros involucrados:

$$X'X = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}, \; X'\mathbf{Y} = \begin{pmatrix} \sum_{j=1}^{10} Y_{1j} \\ \sum_{j=1}^{10} Y_{2j} \end{pmatrix} = \begin{pmatrix} 50.32 \\ 46.61 \end{pmatrix}$$

Por lo tanto, las ecuaciones normales toman la forma:

$$\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} \hat{\mu_1} \\ \hat{\mu_2} \end{pmatrix} = \begin{pmatrix} 50.32 \\ 46.61 \end{pmatrix}$$

Finalmente, $\hat{\mu_1} = 5.032$ y $\hat{\mu_2} = 4.661$. Es decir, el estimador para el peso medio de las plantas en el grupo control es el promedio de los pesos observados en dicho grupo, 5.032. Otro tanto sucede para el tratamiento; en este caso, el promedio de los pesos observados es 4.661.

Una vez resuelto el problema de la estimación de los parámetros, aún debemos determinar si la diferencia existente entre ambos promedios (5.032 vs 4.661) es lo suficientemente grande como para permitirnos afirmar que los ambientes (control y enriquecido) influyen en el crecimiento de las plantas, o puede deberse tan sólo al azar. La solución a este tipo de pregunta será considerada en el siguiente capítulo.

4.2.1 Modelo de Regresión Lineal

En general, se denomina *Modelo de Regresión Lineal* a un modelo de la forma

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i,$$

donde $x_{1i}, \ldots, x_{ki}$ son los valores de k variables independientes.

El caso más sencillo de este tipo de modelos es el *Modelo de Regresión Lineal Simple*, así llamado porque involucra una sola variable independiente x

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

La forma vectorial de este modelo es

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

De aquí se deduce la forma de las ecuaciones normales:

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

o, equivalentemente

$$\begin{aligned} n\hat{\beta}_0 + \sum_{i=1}^{n} x_i \hat{\beta}_1 &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i \hat{\beta}_0 + \sum_{i=1}^{n} x_i^2 \hat{\beta}_1 &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Al resolver este sistema de ecuaciones obtenemos:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \bar{x}\hat{\beta}_1 \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

Claramente, el ejemplo 4.3 corresponde a este tipo de situación. Al resolver las ecuaciones normales para este ejemplo se obtiene: $\hat{\beta}_0 = 143.82$ y $\hat{\beta}_1 = 1.879$. Por lo tanto, $\hat{y}_i = 143.82 + 1.879 x_i$. Si recordamos que x_i representa el porcentaje de madera presente en la pulpa y que y_i es la resistencia del papel producido, observamos a partir del modelo aquí ajustado que la resistencia parece incrementarse en casi dos unidades por cada 1% de madera que se añade a la pulpa. Poseer una relación de este tipo entre la resistencia y el porcentaje de madera puede resultar muy útil cuando se desea incrementar la resistencia sin incurrir en costos excesivos, es decir, sin añadir un exceso de madera a la pulpa.

A continuación, veamos otro ejemplo para el cual puede plantearse un modelo de regresión, en este caso con dos variables independientes.

Ejemplo 4.4 *Un distibuidor de cervezas está analizando el sistema de entregas de su producto; en particular, está interesado en predecir el tiempo requerido para servir a los detallistas. El ingeniero industrial a cargo del estudio ha sugerido que los factores que influyen sobre el tiempo de entrega son el número de cajas de cerveza y la máxima distancia que debe viajar el despachador. Se tomaron muestras y se obtienen los resultados que se muestran en la tabla 4.4.*

Número de Cajas (X_1)	Distancia (X_2)	Tiempo (Y)
10	30	24
15	25	27
10	40	29
20	18	31
25	22	25
18	31	33
12	26	26
14	34	28
16	29	31
22	37	39
24	20	33
17	25	30
13	27	25
30	23	42
24	33	40

Tabla 4.4: Número de cajas transportadas, distancia recorrida y tiempo de servicio al cliente para 15 muestras de un sistema de reparto de cerveza (Ejemplo 4.4)

En la figura 4.5 se muestran los gráficos del tiempo en función del número de cajas y de la distancia a recorrer.

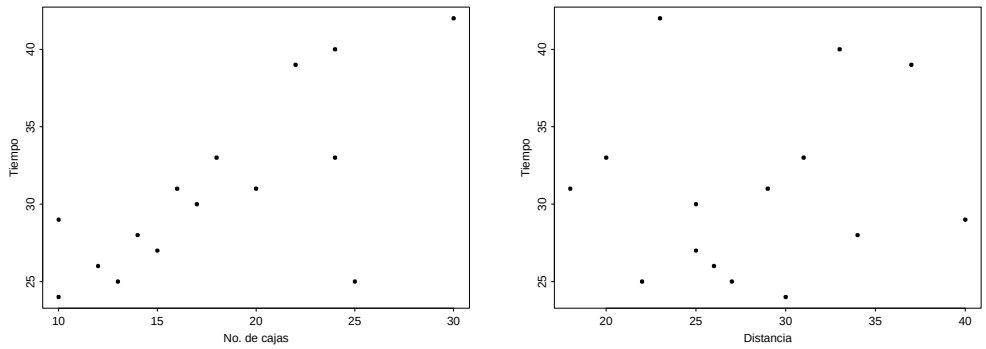

Figura 4.5: Gráficos del tiempo en función del número de cajas y la distancia para el ejemplo de distribución de cerveza.

Supongamos que se decide usar un modelo de la forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.$$

En este caso

$$\mathbf{Y} = \begin{pmatrix} 24 \\ 27 \\ 29 \\ 31 \\ 25 \\ 33 \\ 26 \\ 28 \\ 31 \\ 39 \\ 33 \\ 30 \\ 25 \\ 42 \\ 40 \end{pmatrix}, \; X = \begin{pmatrix} 1 & 10 & 30 \\ 1 & 15 & 25 \\ 1 & 10 & 40 \\ 1 & 20 & 18 \\ 1 & 25 & 22 \\ 1 & 18 & 31 \\ 1 & 12 & 26 \\ 1 & 14 & 34 \\ 1 & 16 & 29 \\ 1 & 22 & 37 \\ 1 & 24 & 20 \\ 1 & 17 & 25 \\ 1 & 13 & 27 \\ 1 & 30 & 23 \\ 1 & 24 & 33 \end{pmatrix}, \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Con lo cual las ecuaciones normales toman la forma

$$\begin{pmatrix} 15 & 270 & 420 \\ 270 & 5364 & 7347 \\ 420 & 7347 & 12308 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 463 \\ 8679 \\ 13027 \end{pmatrix}$$

y al resolverlas se obtiene:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.313 \\ 0.877 \\ 0.456 \end{pmatrix}$$

De aquí que:

$$\hat{y}_i = 2.313 + 0.877x_{1i} + 0.456x_{2i}$$

4.3 Propiedades de los Estimadores de Mínimos Cuadrados

En lo que sigue supondremos que el vector de errores aleatorios en la ecuación (4.3) se distribuye como una normal n-variada con vector de medias $\mathbf{0}$ y matriz de varianza $\sigma^2 I$, es decir,

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I).$$

Por lo tanto,

$$\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$$

Tomando como base estas distribuciones, en esta sección se determinarán algunas propiedades de los estimadores de mínimos cuadrados, las cuales permitirán obtener distribuciones de referencia para probar hipótesis de simplificación para el modelo.

Como $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y} = C\mathbf{Y}$, entonces $\hat{\boldsymbol{\beta}}$ tiene una distribución normal, tal y como se indicó en el apartado 3.2.1. Busquemos entonces su esperanza y su varianza.

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E((X'X)^{-1}X'\mathbf{Y}) \\ &= (X'X)^{-1}X'E(\mathbf{Y}) \\ &= (X'X)^{-1}X'X\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Luego, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, justo el parámetro que queremos estimar. Cuando esto sucede decimos que el estimador, en este caso $\hat{\boldsymbol{\beta}}$, es *insesgado*.

Calculemos $Var(\hat{\boldsymbol{\beta}})$.

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E\{(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))'\} \\ &= E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\} \end{aligned}$$

Como

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (X'X)^{-1}X'\mathbf{Y} - \boldsymbol{\beta} \\ &= (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\ &= (X'X)^{-1}X'\boldsymbol{\varepsilon}, \end{aligned}$$

se obtiene (usando que $X'X$ es simétrica, y por tanto $(X'X)^{-1}$ también lo es)

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E\{((X'X)^{-1}X'\boldsymbol{\varepsilon})((X'X)^{-1}X'\boldsymbol{\varepsilon})'\} \\ &= E\{((X'X)^{-1}X'\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}'X(X'X)^{-1})\} \\ &= (X'X)^{-1}X'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \end{aligned}$$

Finalmente

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}$$

Es decir

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}),$$

y en particular

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}),$$

donde c_{ii} es el elemento de la diagonal de la matriz $C = (X'X)^{-1}$ que corresponde a β_i.

En el caso del ejemplo 4.4, puede calcularse

$$(X'X)^{-1} = \begin{pmatrix} 3.47790 & -0.06857 & -0.07775 \\ -0.06857 & 0.00237 & 0.00092 \\ -0.07775 & 0.00092 & 0.00218 \end{pmatrix}.$$

Entonces, por los resultados anteriores,

$$\begin{aligned}\hat{\beta}_0 &\sim N(\beta_0, 3.4779\sigma^2)\\ \hat{\beta}_1 &\sim N(\beta_1, 0.00237\sigma^2)\\ \hat{\beta}_2 &\sim N(\beta_2, 0.00218\sigma^2)\end{aligned}$$

Ahora bien, no conocemos σ^2, de manera que si queremos obtener algún tipo de inferencia sobre el modelo necesitaremos estimarlo. Para ello, recordemos que

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Si definimos el *vector de residuos* $\mathbf{e}$ como

$$\mathbf{e} = \mathbf{Y} - X\hat{\boldsymbol{\beta}}$$

(es decir, el verdadero valor de $\mathbf{Y}$ menos su valor estimado), resulta razonable pensar que toda la información acerca de la variabilidad aleatoria de los datos está contenida en $\mathbf{e}$, y por lo tanto usaremos éste para estimar σ^2.

Para determinar cómo puede hacerse esta estimación, consideremos previamente algunas propiedades de $\mathbf{e}$.

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - X\hat{\boldsymbol{\beta}}\\ &= \mathbf{Y} - X(X'X)^{-1}X'\mathbf{Y}\end{aligned}$$

Si llamamos $V = X(X'X)^{-1}X'$,

$$\mathbf{e} = (I - V)\mathbf{Y}$$

V es simétrica e idempotente y por tanto $(I-V)$ también lo es. Además, $(I-V)V = 0$. En efecto,

$$\begin{aligned}(I-V)V &= (I - X(X'X)^{-1}X')X(X'X)^{-1}X'\\ &= X(X'X)^{-1}X' - X(X'X)^{-1}X'X(X'X)^{-1}X'\\ &= X(X'X)^{-1}X' - X(X'X)^{-1}X'\\ &= 0\end{aligned}$$

.

$V = X(X'X)^{-1}X'$ es la matriz de la proyección sobre el subespacio generado por las columnas de X, $gen\{X\}$. Por lo tanto, $(I - V)$ es la matriz de proyección sobre el subespacio ortogonal de $gen\{X\}$, $gen\{X\}^{\perp}$.

Luego

$$\begin{aligned} \mathbf{e} &= (I - V)\mathbf{Y} \\ &= (I - V)(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (I - V)\boldsymbol{\varepsilon} \end{aligned}$$

Usando ésto obtenemos la forma cuadrática

$$\begin{aligned} ||\mathbf{e}||^2 = \mathbf{e}'\mathbf{e} &= \mathbf{e}'\mathbf{e} \\ &= \boldsymbol{\varepsilon}'(I - V)'(I - V)\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}'(I - V)\boldsymbol{\varepsilon} \end{aligned}$$

($(I - V)$ es simétrica e idempotente)

Para saber cómo se distribuye esta forma cuadrática, observemos que $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$ y que $(I - V)$ es simétrica e idempotente de rango $n - p$, donde n es el número total de observaciones y p es el número de parámetros del modelo. Entonces, el teorema 3.4 de la sección 3.4 permite afirmar que

$$\frac{1}{\sigma^2}\mathbf{e}'\mathbf{e} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}'(I - V)\boldsymbol{\varepsilon} \sim \chi^2_{n-p}$$

.

En lo que sigue, denotaremos $\mathbf{e}'\mathbf{e} = SSE$ (Suma de Errores al Cuadrado).

Definamos ahora S^2 como

$$S^2 = \frac{(\mathbf{e}'\mathbf{e})}{n - p} = \frac{SSE}{n - p}$$

y observemos que

$$\begin{aligned} E(S^2) &= \frac{1}{n - p}E(\mathbf{e}'\mathbf{e}) \\ &= \frac{\sigma^2}{n - p}E\left(\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}\right) \\ &= \frac{\sigma^2}{n - p}(n - p) \end{aligned}$$

Es decir, $E(S^2) = \sigma^2$. Diremos entonces que S^2 es un *estimador insesgado* para σ^2.

Por último, observemos que

$$SSE = \mathbf{e}'\mathbf{e} = \mathbf{Y}'(I - V)\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}' X'\mathbf{Y}$$

Esta expresión nos permite calcular SSE de una manera más fácil, pues tanto $X'\mathbf{Y}$ como $\hat{\boldsymbol{\beta}}$ han sido calculados al plantear y resolver las exuaciones normales.

Calculemos S^2 para el ejemplo 4.4 (distribución de cerveza). En este caso, $n = 15$, $p = 3$ y por tanto $n - p = 12$. Para el cálculo de SSE,

$$\begin{aligned} SSE &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}' X'\mathbf{Y} \\ &= 14741 - \begin{pmatrix} 2.313 & 0.877 & 0.456 \end{pmatrix} \begin{pmatrix} 463 \\ 8679 \\ 13027 \end{pmatrix} \\ &= 14741 - 14621.802 = 119.198 \end{aligned}$$

Finalmente, podemos calcular S^2

$$S^2 = \frac{SSE}{n - p} = \frac{119.198}{12} = 9.93$$

En el siguiente capítulo veremos cómo usar esta información para realizar inferencia sobre el modelo.