Métodos de la Minería de Datos

Dr. Oldemar Rodríguez Rojas

6 de mayo de 2008

Contents

Elementos básicos de análisis de datos exploratorio

1. Tipos de variables

Definición 1 Una variable describe una característica para el "conjunto de individuos" que fue definida.

Definición 2 El conjunto de individuos es la población en estudio, por ejemplo, el conjunto de todos los costarricenses o las provincias de Costa Rica.

Observación 1 Existen dos tipos de variables: Las cuantitativas y las cualitativas.

- Las variables cuantitativas describen una cantidad (un número real), por ejemplo, el peso de un individuos, el salario o la altura.
- Las variables cualitativas describen una cualidad, por ejemplo el color de los ojos, el diploma que posee la persona. Una variable cualitativa siempre tiene asociada modalidades. Por ejemplo para la variable Y="color de ojos" las modalidades podrían ser: negro, azul, café, otro.

2. Descripción de una variable cuantitativa

Una variable cuantitativa está descrita por los valores que toma en el conjunto de n individuos para los cuales fue definida.

Ejemplo 1

individuo	tamaño
1	1.70
2	1.65
3	1.70
4	1.80

.

Para sintetizar la información contenida en una variable cuantitativa los índices más comunes son:

- El promedio, denotado por $\overline{X}$, que se define por:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

El promedio es un indicador de tendencia central.

- La varianza, denotada por $\text{var}(X)$, definida por:

$$\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2.$$

La varianza es una medida de dispersión que refleja la magnitud de la fluctuación de los datos.

- La desviación estándar, donotada por σ_x, se define por:

$$\sigma_x = \sqrt{\text{var}(X)}.$$

Teorema 1 Si $a, b \in \mathbb{R}$ y X es una variable cuantitativa, entonces:

1. $\overline{aX + b} = a\overline{X} + b$.
2. $\text{var}(aX + b) = a^2\text{var}(X)$.

Prueba.

1.

$$\begin{aligned}
\overline{aX + b} &= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\
&= \left(a \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\right) + b \cdot \frac{1}{n}\sum_{i=1}^{n} 1 \\
&= a\overline{X} + b \cdot \frac{1}{n} \cdot n \\
&= a\overline{X} + b.
\end{aligned}$$

2.

$$\begin{aligned}
\text{var}(aX + b) &= \frac{1}{n}\sum_{i=1}^{n}\left[(ax_i + b) - \overline{(ax + b)}\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - a\overline{x} - b)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\overline{x})^2 \\
&= \frac{1}{n} \cdot a^2\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= a^2\text{var}(X).
\end{aligned}$$

■

Definición 3 Una variable X cuantitativa se dice centrada y reducida si se cumplen:

1. $\overline{X} = 0$.

2. $\text{var}(X) = 1$.

Teorema 2 Sea variable X cuantitativa y Z la variable cuantitativa definida por:

$$z_i := \frac{x_i - \overline{X}}{\sigma_x},$$

entonces la variable Z está centrada y reducida.

Prueba. Se debe probar que la media de Z es cero y que la varianza de Z es uno.

1.

$$\begin{aligned}
\overline{Z} &= \frac{1}{n}\sum_{i=1}^{n} z_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \overline{X}}{\sigma_x} \\
&= \frac{1}{n}\frac{1}{\sigma_x}\sum_{i=1}^{n}(x_i - \overline{X}) \\
&= \frac{1}{n\sigma_x}\left(\sum_{i=1}^{n} x_i - \overline{X}\sum_{i=1}^{n} 1\right) \\
&= \frac{1}{n\sigma_x}\left(n \cdot \frac{1}{n}\sum_{i=1}^{n} x_i - n\overline{X}\right) \\
&= \frac{1}{n\sigma_x}\left(n\overline{X} - n\overline{X}\right) \\
&= 0.
\end{aligned}$$

2.

$$\begin{aligned}
\text{var}(Z) &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{Z})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{\sigma_x} - 0\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{\sigma_x}\right)^2 \\
&= \frac{1}{n\sigma_x^2}\sum_{i=1}^{n}(x_i - \overline{X})^2 \\
&= \frac{1}{\sigma_x^2}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2\right) \\
&= \frac{1}{\text{var}(X)}\text{var}(X) \\
&= 1.
\end{aligned}$$

■

Observación 2 Usualmente cada variable de una tabla de datos se centra y se reduce, de modo que todas las variables tienen la misma media y la misma varianza, evitándose así los efectos de las diferentes escalas.

3. Descripción de una variable cualitativa

Una variable cualitativa también está descrita por los valores que toma en el conjunto de n individuos para los cuales fue definida.

Ejemplo 2

individuo	color de ojos
1	azul
2	verde
3	negro
4	verde
5	azul
6	azul

- La variable "color de ojos" del ejemplo anterior tiene en este caso 3 modalidades: azul, verde y negro.
- Para cada modalidad se define la frecuencia absoluta F, como el número o cantidad de individuos que toma dicha modalidad. Por ejemplo, en

la tabla anterior se tiene que:

$$F(\text{azul}) = 3, F(\text{verde}) = 2 \text{ y que } F(\text{negro}) = 1.$$

- También para cada modalidad se define la frecuencia relativa F_r, que es el cociente entre frecuencia absoluta y el número total de individuos, es decir:
$$F_r = \frac{F}{n}.$$
Por ejemplo, en la tabla anterior se tiene que:
$$F_r(\text{azul}) = \frac{3}{6}, F_r(\text{verde}) = \frac{2}{6} \text{ y que } F_r(\text{negro}) = \frac{1}{6}.$$

- Para facilitar los cálculos en un computador las variables cualitativas se representan usualmente por una matriz $\mathbb{X}$, llamada el código disyuntivo completo, que se define como sigue: Sea X una variable cualitativa con p modalidades definida en un conjunto de n individuos, entonces la matriz $\mathbb{X}$ tendrá tamaño $n \times p$ y su entrada (i, j) es:
$$\mathbb{X}_{ij} = \begin{cases} 1 & \text{si el individuo } i \text{ tomó la modalidad } j \\ 0 & \text{en caso contrario} \end{cases}.$$

Ejemplo 3 Para la variable "color de ojos" del ejemplo 2 el código disyuntivo completo asociado es:

$$\mathbb{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Las columnas de $\mathbb{X}$ se llaman indicatrices. En el ejemplo anterior la columna 1 es la indicatriz de la modalidad azul, la columna 2 es la indicatriz de la modalidad verde y la columna 3 es la indicatriz de la modalidad negro.

Es importante notar que en una fila de la matriz $\mathbb{X}$ nunca habrá más de un 1. Esto garantiza los siguientes dos teoremas.

Teorema 3 La suma de las columnas de la matriz $\mathbb{X}$ es el siguiente vector de tamaño $n \times 1$:

$$\overrightarrow{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Otra propiedad muy útil de $\mathbb{X}$ se expresa en el siguiente teorema.

Teorema 4 $\mathbb{X}^t\mathbb{X}$ es una matriz $p \times p$ diagonal tal que en cada entrada de la diagonal está la frecuencia absoluta de la modalidad respectiva.

Prueba. La entrada $\mathbb{X}_{ij}$ se obtiene efectuando el producto de la fila i de $\mathbb{X}^t$ (columna i de $\mathbb{X}$) por la columna j de $\mathbb{X}$. Si $i \neq j$, dado que nunca existen 2 unos en una fila, el producto de dos columnas columnas diferentes es 0. Si $i = j$, lo que hace el producto de la columna j por si misma, que es un conteo de cuántos individuos tomaron la modalidad j, es decir la frecuencia absoluta de la modalidad j. ■

Ejemplo 4 Para la variable "color de ojos" del ejemplo 2 se tiene que:

$$\mathbb{X}^t\mathbb{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Algunas veces una variable cualitativa se "convierte" en variable cuantitativa asignando un código numérico a cada modalidad, por ejemplo color azul= 1, color verde= 2 y color negro= 3. Así la variable "color de ojos" del ejemplo 2 se puede representar por el vector columna:

$$\text{color de ojos} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 2 \\ 1 \\ 1 \end{pmatrix},$$

nótese que:

$$\begin{pmatrix} 1 \\ 2 \\ 3 \\ 2 \\ 1 \\ 1 \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

Se debe tener cuidado con este tipo de codificación ya que las operaciones algebraicas podrían no tener ningún sentido.

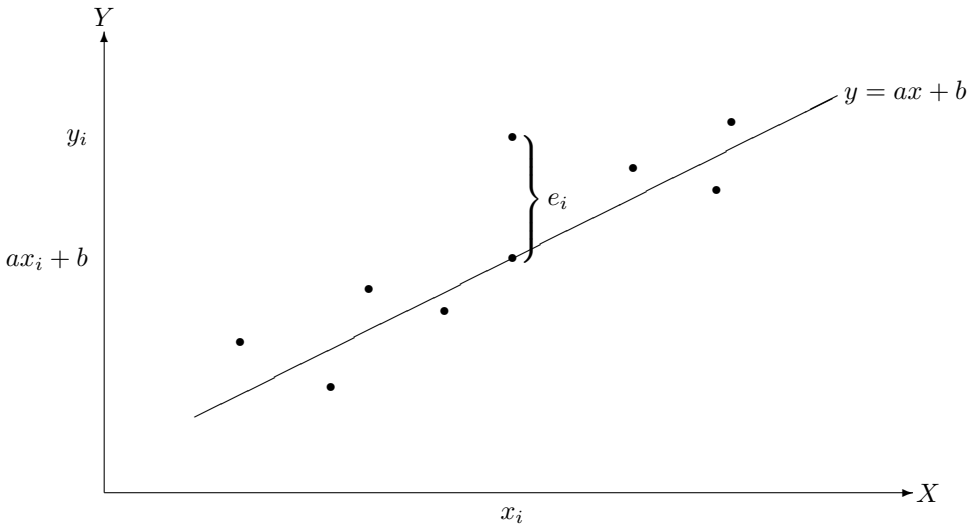

FIGURE 1. Recta de regresión

4. Relación entre dos variables cuantitativas

4.1. La regresión lineal (simple)

Ejemplo 5 Entre las variables "peso" y "altura" usualmente existe una relación, los individuos más pesados en general son los más altos. Esto puede ser cierto, pero no siempre, pues puede haber una persona más pesada que otra y sin embargo su estatura podría ser menor.

Sean X y Y dos variables cualitatativas, la idea es determinar si existe una relación (aproximada) lineal entre X y Y como se muestra en la Figura 1. El problema matemático es encontrar $a, b \in \mathbb{R}$ tal que:

$$y_i = ax_i + b + e_i \text{ para } i = 1, 2, \ldots, n$$

de modo que los e_i sean los más pequeños posibles. Concretamente se usa el criterio de los mínimos cuadrados, es decir, se quiere que la suma $\sum_{i=1}^{n} e_i^2$ sea mínima.

Teorema 5 La recta de regresión de mínimos cuadrados entre dos variables cuantitativas X y Y (i.e. que minimiza $\sum_{i=1}^{n} e_i^2$) se calcula como sigue:

$$\begin{aligned} a &= \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n\text{var}(X)} \\ b &= \overline{Y} - a\overline{X} \end{aligned}$$

Además:

$$\sum_{i=1}^{n} e_i = 0.$$

Prueba. Sean $a, b \in \mathbb{R}$ tal que $y_i = ax_i + b + e_i$ y la suma $\sum_{i=1}^{n} e_i^2$ sea mínima.

- Probaremos primero que $\sum_{i=1}^{n} e_i = 0$.

  Para esto supongamos por contradicción que $\frac{1}{n}\sum_{i=1}^{n} e_i = c$ con $c \neq 0$.

  Note que esto es equivalente a $\sum_{i=1}^{n} e_i = nc$.

  Nótese que $y_i = ax_i + b + e_i = ax_i + (b + c) + (e_i - c)$. Pero:

$$\begin{aligned} \sum_{i=1}^{n}(e_i - c)^2 &= \sum_{i=1}^{n}(e_i^2 - 2e_i c + c^2) \\ &= \sum_{i=1}^{n} e_i^2 - 2c\sum_{i=1}^{n} e_i + c^2\sum_{i=1}^{n} 1 \\ &= \sum_{i=1}^{n} e_i^2 - 2cnc + c^2 n \\ &= \sum_{i=1}^{n} e_i^2 - 2c^2 n + c^2 n \\ &= \left(\sum_{i=1}^{n} e_i^2\right) - c^2 n \\ &< \sum_{i=1}^{n} e_i^2 \end{aligned}$$

Esto implica que $\sum_{i=1}^{n}(e_i - c)^2 < \sum_{i=1}^{n} e_i^2$, lo cual es una contradicción ya que $\sum_{i=1}^{n} e_i^2$ es mínimo.

Por lo tanto $c = 0$ lo que implica que $\sum_{i=1}^{n} e_i = 0$.

- Probaremos que: $b = \overline{Y} - a\overline{X}$.

  Note que $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n} \cdot 0 = 0$ lo que implica que:

$$\begin{aligned} \overline{Y} &= \overline{ax + b + e} \\ &= a\overline{X} + b + \overline{e} \\ &= a\overline{X} + b, \end{aligned}$$

  de donde $b = \overline{Y} - a\overline{X}$.

- Probaremos que: $a = \dfrac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{n\text{var}(X)}$.

  Como $y_i = ax_i + b + e_i$ entonces:

$$\begin{aligned} e_i &= y_i - ax_i - b \\ &= y_i - ax_i - \overline{Y} + a\overline{X} \\ &= (y_i - \overline{Y}) - a(x_i - \overline{X}) \end{aligned} \tag{1.1}$$

  Esto implica que minimizar $\sum_{i=1}^{n} e_i^2$ es equivalente a minimizar $\sum_{i=1}^{n} \left[(y_i - \overline{Y}) - a(x_i - \overline{X})\right]^2$.

  Para calcular este mínimo se debe derivar $\sum_{i=1}^{n} \left[(y_i - \overline{Y}) - a(x_i - \overline{X})\right]^2$ respecto al único parámetro a e igualar a 0.

$$\begin{aligned} \frac{d}{da}\left(\sum_{i=1}^{n} \left[(y_i - \overline{Y}) - a(x_i - \overline{X})\right]^2\right) &= \sum_{i=1}^{n} 2\left[(y_i - \overline{Y}) - a(x_i - \overline{X})\right]\left[-(x_i - \overline{X})\right] \\ &= \sum_{i=1}^{n} \left[-2(y_i - \overline{Y})(x_i - \overline{X}) + 2a(x_i - \overline{X})^2\right] \\ &= -2\sum_{i=1}^{n}(y_i - \overline{Y})(x_i - \overline{X}) + 2a\sum_{i=1}^{n}(x_i - \overline{X})^2 \end{aligned}$$

Así:

$$\frac{d}{da}\left(\sum_{i=1}^{n}\left[(y_i-\overline{Y})-a(x_i-\overline{X})\right]^2\right)=0$$

es equivalente a:

$$-2\sum_{i=1}^{n}(y_i-\overline{Y})(x_i-\overline{X})+2a\sum_{i=1}^{n}(x_i-\overline{X})^2=0,$$

de donde:

$$a=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{\sum_{i=1}^{n}(x_i-\overline{X})^2},$$

es decir:

$$a=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n\text{var}(X)}.$$

■

Definición 4 Se define la covarianza entre dos variables cualitativas X y Y como sigue:

$$\text{cov}(X,Y):=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y}).$$

Teorema 6 La pendiente de la recta de regresión se puede calcular como sigue:

$$a=\frac{\text{cov}(X,Y)}{\text{var}(X)}.$$

Prueba. Es claro de la definición 4. ■

4.2. La ecuación de análisis de la varianza

Teorema 7 En el modelo $y_i=ax_i+b+e_i$ se tiene que:

$$\text{var}(Y)=\text{var}(aX+b)+\text{var}(e)$$

donde:

- var(Y) es la varianza de Y.
- var$(aX+b)$ es la varianza explicada por las variaciones de X.

- $\text{var}(e)$ es la varianza de los residuos.

Prueba. De (1.1) se sabe que $e_i = (y_i - \overline{Y}) - a(x_i - \overline{X})$, de donde:

$$y_i - \overline{Y} = a(x_i - \overline{X}) + e_i.$$

Esto implica que:

$$\begin{aligned}
\text{var}(Y) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[a(x_i - \overline{X}) + e_i]^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} a^2(x_i - \overline{X})^2 + 2a\sum_{i=1}^{n}(x_i - \overline{X})e_i + \sum_{i=1}^{n} e_i^2\right)
\end{aligned}$$

Se sabe que:

$$\text{var}(aX + b) = a^2\text{var}(X) = \frac{a^2}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2,$$

además se sabe que:

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \text{var}(e).$$

Basta probar que:

$$\sum_{i=1}^{n}(x_i - \overline{X})e_i = 0.$$

Vamos esto:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})e_i &= \operatorname{cov}(X,e)\\
&= \operatorname{cov}(X,Y-aX-b)\\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-ax_i-b-\overline{Y}+a\overline{X}+b)\\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y}-a(x_i-\overline{X}))\\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})-a\sum_{i=1}^{n}(x_i-\overline{X})^2\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})-\frac{a}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2\\
&= \operatorname{cov}(X,Y)-a\operatorname{var}(X)\ \left(\text{pues } a=\frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}\right)\\
&= \operatorname{cov}(X,Y)-\operatorname{cov}(X,Y)\\
&= 0.
\end{aligned}$$

■

Definición 5 El **coeficiente de determinación** se define por:

$$R^2=\frac{\operatorname{var}(aX+b)}{\operatorname{var}(Y)}=\frac{\text{varianza explicada por la recta}}{\text{varianza total}}.$$

Teorema 8

$$R=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Prueba.

$$\begin{aligned} R^2(X,Y) &= \frac{\text{var}(aX+b)}{\text{var}(Y)} \\ &= \frac{a^2\text{var}(X)}{\text{var}(Y)} \\ &= \frac{\text{cov}^2(X,Y)\text{var}(X)}{\text{var}^2(X)\text{var}(Y)} \\ &= \frac{\text{cov}^2(X,Y)}{\text{var}(X)\text{var}(Y)} \\ &= \frac{\text{cov}^2(X,Y)}{\sigma_X^2\sigma_Y^2} \\ &= \left(\frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}\right)^2. \end{aligned}$$

■

Definición 6 R se llama el **coeficiente de correlación**.

4.3. Interpretación geométrica del coeficiente de correlación

- Una variable X que toma n valores puede ser representada como un vector de $\mathbb{R}^n$ (En adelante $\mathbb{R}^n$ se denominará espacio vectorial de las variables).
- En $\mathbb{R}^n$ el producto escalar (producto interno) usual es: Si $X = (x_1, x_2, \ldots, x_n)$ y $Y = (y_1, y_2, \ldots, y_n)$ entonces:

$$\langle X, Y\rangle = \sum_{i=1}^{n} x_i y_i.$$

- En análisis de datos se usa el siguiente producto interno:

$$\langle X, Y\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

- Con este producto interno se tiene la interpretación del coeficiente de correlación que se expone en el siguiente teorema.

Teorema 9 En el espacio vectorial de las variables $\mathbb{R}^n$ el coseno del ángulo entre 2 variables centradas y reducidas es igual al coeficiente de correlación entre esas 2 variables.

Corolario 1

$$\|X\| = \sigma_X.$$

5. Regresión múltiple

PROBLEMA: Se trata de explicar (predecir) una variable cuantitativa Y, observada en n individuos, utilizando m variables cuantitativas linealmente independientes X_j con $j = 1, 2, \ldots, m$. Estas variables también han sido observadas en los mismos n individuos.

Ejemplo 6 Consideremos que los principales factores que inciden en el rendimiento del cultivo del trigo son: a) Potasio y fósforo (kg/Ha), b) Nitrógeno (kg/Ha), c) Agua de lluvia promedio (cm^3), d) Acidez del suelo (pH) y e) Temperatura promedio (°). Y que las respectivas variables explicativas se denotarán como:

X_1: Cantidad de potasio y fósforo (mezcla) por hectárea.
X_2: Cantidad de nitrógeno por hectárea.
X_3: Acidez del suelo (pH).
X_4: Cantidad promedio de lluvia caída (cm^3).
X_5: Temperatura promedio.
Y: Rendimiento medido en quintales/Ha. (variable dependiente a explicar)

	X_1	X_2	X_3	X_4	X_5	Y
	Pot	Nit	A.lluv	PH	tem	rend
1	1100	300	6	5	10	30
2	1000	200	4	7	8	20
3	1200	350	6.7	8	10	40
4	1000	300	5	6	8	25
5	1100	300	5.5	7	9	35
6	1200	350	8	6	11	45
7	900	300	4	5	8	30
8	700	400	3.5	3	7	25
9	1200	350	6	7	7	35
10	1300	350	7	6.5	10	40

TABLE 1.1. Rendimiento de trigo, Y, en 10 parcelas según los factores X_i.

Nuestro objetivo es medir de alguna forma la incidencia en la variable Y de los factores explicativos. Asumimos que el efecto de cada factor sobre el rendimiento es aditivo de manera que Y se expresa en términos de los factores explicativos como:

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_5X_5 + e. \tag{1.2}$$

Al conocer los valores de los a_i de este modelo, se explica el poder de incidencia de cada factor X_i en la variable Y, lo que permite estimar el

rendimiento del cultivo de trigo Y, para diferentes valores de los factores explicativos X_i (variables independientes), entre otros objetivos.

A fin de disponer de los datos necesarios para estimar los valores $a_0, \ldots, a_5$ se realizó el siguiente experimento: cultivar trigo en 10 campos diferentes sometido a distintos tratamientos de las variables explicativas. Las mediciones resultantes de las 6 variables, en cada caso, se resumen en la tabla 1.1.

- Es una generalización de la regresión lineal, es decir:

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_m x_{im} + e_i \tag{1.3}$$

donde:

- y_i es la observación $i-$ésima de la variable y (es decir, en el individuo i.)
- x_{ij} es la observación $i-$ésima de la variable x_j (es decir, en el individuo i.)
- Los coeficientes a_j para $j = 1, 2, \ldots, m$ son las incognitas y serán calculados usando el criterio de mínimos cuadrados, es decir de modo tal que $\sum_{i=1}^{n} e_i^2$ sea mínimo.

- Análogamente al caso lineal se puede probar que $\sum_{i=1}^{n} e_i = 0$ (tarea).

$$\Rightarrow \bar{e} = 0,$$

$$\Rightarrow \bar{y} = \overline{a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m + e_i},$$

$$\Rightarrow \bar{y} = a_0 + a_1 \bar{x}_1 + a_2 \bar{x}_2 + \cdots + a_m \bar{x}_m. \tag{1.4}$$

Restando (1.3)$-$(1.4) se tiene que:

$$y_i - \bar{y} = a_1(x_{i1} - \bar{x}_1) + a_2(x_{i2} - \bar{x}_2) + \cdots + a_m(x_{im} - \bar{x}_m) + e_i. \tag{1.5}$$

Si denotamos por Y_i y X_j las variables centradas:

$$Y_i = y_i - \bar{y} \quad \text{y} \quad X_{ij} = x_{ij} - \bar{x}_j \ \text{ para } \ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m,$$

entonces (1.5) se escribe como:

$$Y_i = a_1 X_{i1} + a_2 X_{i2} + \cdots + a_m X_{im} + e_i, \tag{1.6}$$

para $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$. Luego matricialmente (1.5) se escribe como:

$$Y = Xa + e, \tag{1.7}$$

donde:

$$X = \begin{pmatrix} X_{11} & \cdots & X_{1j} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & \cdots & X_{ij} & \cdots & X_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nj} & \cdots & X_{nm} \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{pmatrix},$$

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_m \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{pmatrix}.$$

5.1. Interpretación en el espacio de los individuos

- El espacio de los individuos es el espacio filas de la matriz por bloques $[X|Y]$, es decir se tienen n puntos en $\mathbb{R}^{m+1}$.
- Minimizar $\sum_{i=1}^{n} e_i^2$ corresponde a encontrar un hiperplano:

$$Y_i = a_1 X_1 + a_2 X_2 + \cdots + a_m X_m$$

en $\mathbb{R}^{m+1}$ de tal manera que pase lo más cerca posible de los n puntos (individuos).

5.2. Interpretación en el espacio de las variables

En espacio de las variables $\mathbb{R}^n$ están las columnas de X, denotadas por $X_1, X_2, \ldots, X_m$. También están en este espacio los vectores Y y e. La interpretación está dada por el siguiente Teorema:

Teorema 10 Encontrar los parámetros $a_1, a_2, \ldots, a_m$ de tal manera que la $\sum_{i=1}^{n} e_i^2$ sea mínima es equivalente a proyectar ortogonalmente el vector Y sobre el espacio generado por $X_1, X_2, \ldots, X_m$.

$$\mathcal{B} = \{X_1, X_2, \ldots, X_m\} \mapsto \{v_1, v_2, \ldots, v_m\} \text{ ortonormal}$$

$$Y \approx Xa = \langle Y, v_1\rangle v_1 + \langle Y, v_2\rangle v_2 + \cdots + \langle Y, v_m\rangle v_m$$

$$Xa = X(X^tX)^{-1}X^ty$$

Corolario 2 El vector que minimiza la suma del cuadrado de los residuos es:

$$a = (X^tX)^{-1}X^ty.$$

Observación 3 Para pronosticar una variable Y, a partir de una tabla de datos centrada X, se utiliza:

$$Y \approx \tilde{Y} = Xa, \tag{1.8}$$

Observación 4 Para pronosticar una variable Y, a partir de una tabla de datos **NO** centrada X, se utiliza:

$$Y \approx \tilde{Y} = Xa + a_0 \cdot 1_n, \tag{1.9}$$

donde:

$$a_0 = \bar{y} - \sum_{j=1}^{m} a_j\bar{x}_j \text{ y } 1_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}$$