Laboratorio de Datos
Modelo Lineal Multivariado

Primer Cuatrimestre 2025
Turnos noche

Facultad de Ciencias Exactas y Naturales, UBA

Modelos cuantitativos

Modelar una variable respuesta («dependiente») $y \in \mathbb{R}$ en función de ciertas variables explicativas («independientes») $(x_1, \ldots, x_d) \in \mathbb{R}^d$, consiste en encontrar una función $\hat{y} = f(x_1, \ldots, x_d)$ que aproxime razonablemente bien los valores reales.

Típicamente fijamos una función f que dependa de ciertos parámetros y elegimos los parámetros que minimicen el error en los datos que tenemos.

Modelos cuantitativos

Ejemplos:

- Una fórmula que relacione el radio con el volumen de una esfera.
- Una fórmula que relacione el peso con la altura de una persona.
- Una fórmula que nos diga el valor de una propiedad según la cantidad de metros cuadrados y barrio donde se ubica el departamento.
- Una fórmula que nos diga la cantidad de calorías de un alimento según la cantidad de proteínas, grasas e hidratos de carbono.
- Una fórmula que nos diga los gastos en tarjetas de crédito de un potencial cliente de un banco en base a edad, sueldo, cantidad de hijos y otras variables del cliente.

Modelos univariados

Hemos visto modelos univariados ($d = 1$) ajustando una función lineal:

$$y = f(x) = \beta_0 + \beta_1 x,$$

y modelos univariados ajustando una función polinomial:

$$y = \text{Poli}_k(x) = \sum_{i=0}^{k} \beta_i x^i = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k.$$

Estos son casos particulares de modelos **lineales** (aunque usted no lo crea!).

Repaso: sistemas de ecuaciones lineales

Los modelos multivariados son una extensión directa de los univariados, la mayor diferencia es que no podemos graficar los resultados en un plano.

Ejemplo. Sabemos que un fondo de inversión invirtió en acciones de YPF, Santander y Nvidia (y solo en estas acciones) pero no sabemos cuántas acciones compró de cada una. ¿Cómo podemos averiguarlo?

Repaso: sistemas de ecuaciones lineales

Los modelos multivariados son una extensión directa de los univariados, la mayor diferencia es que no podemos graficar los resultados en un plano.

Ejemplo. Sabemos que un fondo de inversión invirtió en acciones de YPF, Santander y Nvidia (y solo en estas acciones) pero no sabemos cuántas acciones compró de cada una. ¿Cómo podemos averiguarlo?

Suponemos que tenemos disponible:

- La valorización del fondo al final de cada día.
- El valor de la acción de cada empresa al cierre de cada día.

Sistemas de ecuaciones lineales

Ponemos toda la información en la siguiente tabla.

Total	YPF	Santander	Nvidia
170262.00	20935	20100	37100.0
169929.50	21030	20500	36255.0
171064.00	20770	21700	36000.0
169637.35	20950	21000	35645.5
164625.45	20750	20316	33878.5

Table: Valores diarios de acciones

Planteamos el sistema lineal

Llamamos c_1, c_2 y c_3 a la cantidad de acciones de cada tipo. Para calcular los valores, tenemos que resolver el siguiente sistema de ecuaciones:

$$
\begin{array}{lcccc}
 & & YPF & Santander & Nvidia \\
 & & \downarrow & \downarrow & \downarrow \\
\text{Día 1} \rightarrow & 170262.00 = & 20935c_1 & + 20100c_2 & + 37100.0c_3 \\
\text{Día 2} \rightarrow & 169929.50 = & 21030c_1 & + 20500c_2 & + 36255.0c_3 \\
\text{Día 3} \rightarrow & 171064.00 = & 20770c_1 & + 21700c_2 & + 36000.0c_3 \\
\text{Día 4} \rightarrow & 169637.35 = & 20950c_1 & + 21000c_2 & + 35645.5c_3 \\
\text{Día 5} \rightarrow & 164625.45 = & 20750c_1 & + 20316c_2 & + 33878.5c_3
\end{array}
$$

Resolvemos el sistema

Como tenemos 3 incógnitas, nos alcanza con 3 ecuaciones:

$$170262.00 = 20935c_1 + 20100c_2 + 37100.0c_3$$
$$169929.50 = 21030c_1 + 20500c_2 + 36255.0c_3$$
$$171064.00 = 20770c_1 + 21700c_2 + 36000.0c_3$$

Para resolver el sistema, construimos la matriz ampliada

$$\left(\begin{array}{ccc|c} 20935.0 & 20100.0 & 37100.0 & 170262.00 \\ 21030.0 & 20500.0 & 36255.0 & 169929.50 \\ 20770.0 & 21700.0 & 36000.0 & 171064.00 \end{array}\right)$$

¿Qué hay en las primeras 3 columnas de la matriz? ¿Qué hay en la última columna?

Solución del sistema

Triangulando la matriz y despejando, obtenemos los valores

$$c_1 = 3.2, \qquad c_2 = 2.0, \qquad c_3 = 1.7.$$

Estas son las cantidades de cada acción que tiene el fondo de inversión.

Podemos verificar fácilmente que estos valores satisfacen también las otras dos ecuaciones.

Notación matricial

Podemos escribir el sistema de ecuaciones en forma compacta usando notación matricial:

$$\begin{pmatrix} 20935.0 & 20100.0 & 37100.0 \\ 21030.0 & 20500.0 & 36255.0 \\ 20770.0 & 21700.0 & 36000.0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 170262.00 \\ 169929.50 \\ 171064.00 \end{pmatrix}$$

Obtenemos un sistema de la forma

$$Xc = y$$

(en álgebra lineal sería más común $Ax = b$).

¿Qué hay en las columnas de X? ¿Que hay en la matriz y?

Modelo lineal multivariado

El caso recién visto fue un ejemplo de juguete, donde existe una relación lineal entre las variables que tenemos que estimar.

En las aplicaciones reales, esa formula puede no existir (por ejemplo, si queremos estimar los gastos en tarjeta de crédito de una persona en función del sueldo y la cantidad de hijos), pero el modelo lineal puede darnos una buena estimación.

En el modelo lineal suponemos que nuestros datos satisfacen una relación del tipo

$$Xc = y,$$

donde y es la variable que queremos explicar y X es una matriz que podemos construir a partir de los demás datos.

Como esa relación en general no existe, buscamos un vector c que haga que Xc se parezca lo más posible a y.

Más ecuaciones que variables - La vida real

Cuando consideramos un sistema con más ecuaciones que variables, en general **NO** tiene solución.

Incluso si teóricamente existe solución, en la práctica siempre aparecen errores numéricos y no podemos determinar si un sistema tiene solución (numéricamente es MUY difícil saber si un número es igual a 0 o no).

Solución: en vez de buscar una solución exacta del sistema de ecuaciones

$$Xc = y,$$

buscamos un vector c que minimice el error, es decir, que haga pequeñas las coordenadas del vector de errores

$$Xc - y.$$

El milagro de los mínimos cuadrados

Llegamos así al método de mínimos cuadrados. El vector c que minimiza la suma de los errores al cuadrado del sistema

$$Xc = y,$$

es solución del sistema lineal de ecuaciones

$$X^T X c = X^T y.$$

Es un sistema cuadrado y en general tiene solución única.

El problema DIFÍCIL de minimizar los errores se transforma en el problema FÁCIL de resolver un sistema lineal de ecuaciones. **Este es el milagro de los mínimos cuadrados.**

Ejercicio: modelos lineales

Dadas una variable a predecir y y variables explicativas $x_1, x_2, \ldots$, ¿cuáles de los siguientes modelos son lineales? ¿Cuál es la matriz X en cada caso?

1. $y = c_1 x_1 + c_2 x_2$
2. $y = c_0 + c_1 x_1 + c_2 x_2$
3. $y = c_0 + c_1 x_1 + c_2 x_1^2$
4. $y = c_0 + c_1 x_1 + x_1^{c_2}$
5. $y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_1 x_2$
6. $y = c_0 + c_1 \sin(x_1) + c_2 \sin(x_2)$
7. $y = c_0 + c_1 \sin(c_2 + x_1)$
8. $y = c_0 + c_1 e^{x_1}$
9. $y = c_0 \cdot c_1^{x_1}$

Ejercicio: modelos lineales

Dadas una variable a predecir y y variables explicativas $x_1, x_2, \ldots$, ¿cuáles de los siguientes modelos son lineales? ¿Cuál es la matriz X en cada caso?

1. $y = c_1 x_1 + c_2 x_2$
2. $y = c_0 + c_1 x_1 + c_2 x_2$
3. $y = c_0 + c_1 x_1 + c_2 x_1^2$
4. $y = c_0 + c_1 x_1 + x_1^{c_2}$
5. $y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_1 x_2$
6. $y = c_0 + c_1 \sin(x_1) + c_2 \sin(x_2)$
7. $y = c_0 + c_1 \sin(c_2 + x_1)$
8. $y = c_0 + c_1 e^{x_1}$
9. $y = c_0 \cdot c_1^{x_1}$

Algunos de estos modelos se pueden linearizar, pero eso ya es otra historia...

Ejemplo real: cálculo de calorías

Vamos a construir un modelo para un caso real: calcular las calorías de un alimento en función de sus componentes, utilizando las herramientas de Python para la construcción de modelos.

1. Proponemos un modelo apropiado.
2. Construimos las matrices X e y utilizando `Formulaic`.
3. Ajustamos el modelo utilizando `linear_model.fit()`.
4. Calculamos el error en el ajuste.