Laboratorio de Datos
Entrenamiento y testeo

Primer Cuatrimestre 2024
Turnos tarde y noche

Facultad de Ciencias Exactas y Naturales, UBA

Introducción: sistemas de ecuaciones lineales

Sabemos que un fonde 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?

Introducción: sistemas de ecuaciones lineales

Sabemos que un fonde 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 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

Cuadro: 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:

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

Resolvemos el sistema

Como tenemos 3 incógnitas, nos alcanza con 3 ecuaciones:

$$\begin{aligned} 170262.00 &= 20935c_1 + 20100c_2 + 37100.0c_3 \\ 169929.50 &= 21030c_1 + 20500c_2 + 36255.0c_3 \\ 171064.00 &= 20770c_1 + 21700c_2 + 36000.0c_3 \end{aligned}$$

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.

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

$$Xc = y$$

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

Verificación del "modelo"

En base a los resultados que obtuvimos, ¿podemos confirmar que las acciones del fondo son las 3 acciones que usamos?

Si no estamos seguros si eran acciones de Santander o Galicia, ¿cómo podríamos asegurarnos?

Cambiamos los valores de Santander por los valores de Galicia:

$$
\begin{aligned}
170262.00 &= 20935c_1 + \mathbf{20100}c_2 + 37100.0c_3 \\
169929.50 &= 21030c_1 + \mathbf{19400}c_2 + 36255.0c_3 \\
171064.00 &= 20770c_1 + \mathbf{21900}c_2 + 36000.0c_3
\end{aligned}
$$

Sobreajuste

Resolvemos el sistema y obtenemos estos valores:

$$c_1 = 6.69507872, \qquad c_2 = 1.16828332, \qquad c_3 = 0.17838362$$

Los números no se ven menos redondos, pero eso no alcanza para decidir cuál es el modelo correcto.

Un sistema de 3 ecuaciones y 3 incógnitas en general **siempre** tiene solución.

Sobreajuste (overfitting). Cuando tenemos igual cantidad de parámetros que observaciones, el sistema (casi) siempre va a tener solución pero no nos da ninguna información sobre si el modelo es correcto, no podemos usarlo para estimar otros valores.

Selección de modelos

¿Qué estrategias se les ocurren para ver cuál modelo es mejor?

Selección de modelos

¿Qué estrategias se les ocurren para ver cuál modelo es mejor?

1. Verificar la fórmula en otros días.

Selección de modelos

¿Qué estrategias se les ocurren para ver cuál modelo es mejor?

1. Verificar la fórmula en otros días.
2. Utilizar más días al plantear el sistema de ecuaciones.

Selección de modelos

¿Qué estrategias se les ocurren para ver cuál modelo es mejor?

1. Verificar la fórmula en otros días.
2. Utilizar más días al plantear el sistema de ecuaciones.

Las dos estretegias son ideas centrales en la construcción de modelos:

1. Probar el modelo en datos distintos a los que usamos para construir el modelo.
2. Utilizar la mayor cantidad posible de datos para construir el modelo.

1. Conjuntos de entrenamiento y testeo

Separamos nuestro conjunto de datos en dos subconjuntos:

- **Conjunto de entrenamiento.** Lo utilizamos para construir el modelo. En un modelo lineal, lo usamos para calcular los coeficientes (c_1, c_2, c_3).
- **Conjunto de testeo.** Lo utilizamos para verificar si el modelo construido ajusta bien a los datos en este conjunto.

2. Más ecuaciones que variables - Ejemplo de juguete

Si consideramos el sistema original, tenemos 5 ecuaciones y 3 variables.

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

En este ejemplo (de juguete) si utilizamos los datos de Santander, el sistema tiene solución. Si usamos los datos de Galicia el sistema no tiene solución.

2. Más ecuaciones que variables - La vida real

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

Aunque teóricamente exista 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 DIFICIL 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.**

Resumen

Tenemos dos modelos posibles y queremos elegir el más apropiado:

$$A) total = c_1 \cdot YPF + c_2 \cdot Santander + c_3 \cdot Nvidia$$
$$B) total = c_1 \cdot YPF + c_2 \cdot Galicia + c_3 \cdot Nvidia$$

Seguimos los siguientes pasos:

1. Buscamos datos de la mayor cantidad posible de días.
2. Separamos el conjunto en dos: conjunto de entrenamiento ($80\,\%$ de los días) y conjunto de testeo ($20\,\%$ restante)
3. Calculamos c_1, c_2, c_3 para cada uno de los dos modelos utilizando mínimos cuadrados en el conjunto de entrenamiento.
4. Calculamos el error cuadrático medio de lás fórmulas resultantes aplicadas al conjunto de testeo.

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 existe una relación del tipo

$$Xc = y.$$

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

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

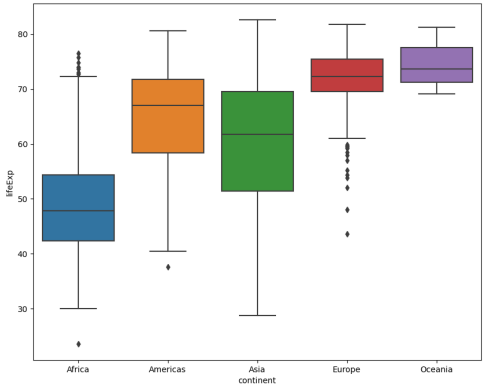
Ejemplo real: cálculo de calorías

Vamos a repetir estos pasos en un caso real: calculas las calorías de un alimento en función de sus componentes, utilizando las herramientas de Python para la construcción de modelos.

1. Construimos las matrices X e y utilizando `Formulaic`.
2. Separamos las matrices en entrenamiento y testeo utilizando `train_test_split`
3. Ajustamos el modelo utilizando `linear_model.fit()`.
4. Calculamos el error en los datos de testeo utilizando `linear_model.predict()`.

Box plot

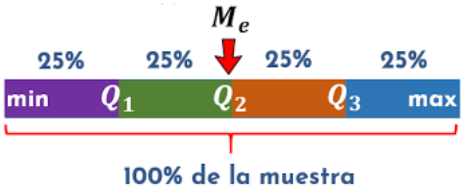
Un gráfico **box plot** usa cajas y líneas para mostrar información de la distribución de uno o más grupos de datos numéricos.



Elementos de un Box plot: cuartiles

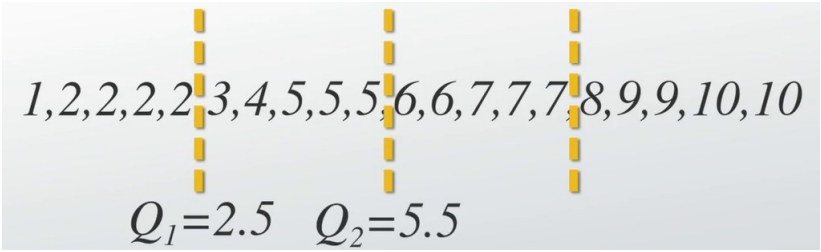
Dada una variable numérica, ordenamos los valores y los partimos en 4 grupos de igual tamaño.

- Primer cuartil (Q1): es un valor mayor que el $25\,\%$ de los datos y menor que el otro $75\,\%$.
- Segundo cuartil (Q2): es un valor mayor que el $50\,\%$ de los datos y menor que el otro $50\,\%$ (corresponde a la mediana).
- Tercer cuartil (Q3): es un valor mayor que el $75\,\%$ de los datos y menor que el otro $25\,\%$.



Cuartiles: ejemplo

Ordenamos los datos de menor a mayor y tomamos valores que dividen a los datos en 4 partes iguales.



Cálculo de cuartiles en Wikipedia...

Métodos de computación [editar]

Distribuciones discretas [editar]

Para distribuciones discretas, no existe un acuerdo universal sobre la selección de los valores de los cuartiles. [2]

Método 1 [editar]

1. Utilice la mediana para dividir el conjunto de datos ordenados en dos mitades. La mediana se convierte en los segundos cuartiles.
   - Si hay un número impar de puntos de datos en el conjunto de datos ordenados original, **incluya** la mediana (el valor central en la lista ordenada) en ninguna de las mitades.
   - Si hay un número par de puntos de datos en el conjunto de datos ordenado original, divida este conjunto de datos exactamente por la mitad.
2. El valor del cuartil inferior es la mediana de la mitad inferior de los datos. El valor del cuartil superior es la mediana de la mitad superior de los datos.

Esta regla la emplean el diagrama de caja de la calculadora TI-83 y las funciones "Estadísticas de 1 var".

Método 2 [editar]

1. Utilice la mediana para dividir el conjunto de datos ordenados en dos mitades. La mediana se convierte en los segundos cuartiles.
   - Si hay un número impar de puntos de datos en el conjunto de datos ordenados original, **incluya** la mediana (el valor central en la lista ordenada) en ambas mitades.
   - Si hay un número par de puntos de datos en el conjunto de datos ordenado original, divida este conjunto de datos exactamente por la mitad.
2. El valor del cuartil inferior es la mediana de la mitad inferior de los datos. El valor del cuartil superior es la mediana de la mitad superior de los datos.

Los valores encontrados por este método también se conocen como " bisagras de Tukey ", [4] ver también bisagra media .

Método 3 [editar]

1. Utilice la mediana para dividir el conjunto de datos ordenados en dos mitades. La mediana se convierte en los segundos cuartiles.
   1. Si hay un número impar de puntos de datos, vaya al siguiente paso.
   2. Si hay números pares de puntos de datos, entonces el Método 3 comienza igual que el Método 1 o el Método 2 anteriores y puede optar por incluir o no la mediana como un nuevo punto de datos. Si elige incluir la mediana como el nuevo punto de datos, continúe con el paso 2 o 3 a continuación porque ahora tiene un número impar de puntos de datos. Si no elige la mediana como nuevo punto de datos, continúe con el Método 1 o 2 donde comenzó.
2. Si hay ($4\,n$ +1) puntos de datos, entonces el cuartil inferior es el 25% del *enésimo* valor de datos más el 75% del (n +1)ésimo valor de datos; el cuartil superior es el 75 % del punto de datos ($3\,n$ +1) más el 25 % del punto de datos ($3\,n$ +2).
3. Si hay ($4\,n$ +3) puntos de datos, entonces el cuartil inferior es el 75% del (n +1)ésimo valor de datos más el 25% del (n +2)ésimo valor de datos; el cuartil superior es el 25 % del punto de datos ($3\,n$ +2) más el 75 % del punto de datos ($3\,n$ +3).

Método 4 [editar]

Si tenemos un conjunto de datos ordenados$x_1, x_2, \ldots, x_n$, entonces podemos interpolar entre puntos de datos para encontrar elp^{th} cuantil empírico six_i está en el$i/(n+1)$cuantil. Si denotamos la parte entera de un númeroαpor$\lfloor \alpha \rfloor$, entonces la función cuantil empírica viene dada por,

$$q(p/4) = x_k + \alpha(x_{k+1} - x_k),$$

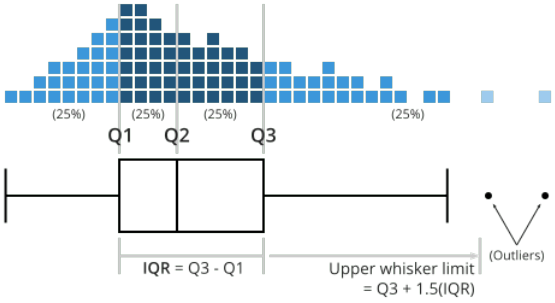
dónde$k = \lfloor p(n+1)/4 \rfloor$y$\alpha = p(n+1)/4 - \lfloor p(n+1)/4 \rfloor$. [1]

Para encontrar el primer, segundo y tercer cuartil del conjunto de datos, evaluaríamos$q(0.25)$.$q(0.5)$.$yq(0.75)$respectivamente.

Elementos de un Box plot

Dada una variable numérica, ordenamos los valores y los partimos en 4 grupos de igual tamaño.

En un box plot, dibujamos una caja, con límites en Q1 y Q3 y una línea central marcando el valor de Q2.



Elementos de un Box plot

La distancia entre Q3 y Q1 se conoce como *rango intercuartil* (IQR) y se utilizan para trazar los "bigotes".

Cada bigote se extiende hasta el valor más lejano de los datos a una distancia menor a 1.5 veces el valor IQR.

Cualquier valor de los datos más allá de esa distancia se considera un dato atípico (outlier) y se marca con un punto.