Métodos Computacionales 2023
Trabajo Práctico 1:
Regresión Lineal Múltiple

1. Definición del problema

En muchos contextos como experimentos y estudios, se busca estudiar el comportamiento de una variable de interés a la que llamamos variable de *respuesta* o *dependiente* y notamos con la letra y. Para ello, se cuenta con un número de variables cuyo comportamiento es conocido, a las cuales llamamos variables *explicativas* o *independientes* y denotamos con $x_1, x_2, ..., x_p$.

Un enfoque posible es intentar dar una representación simplificada que capture la relación entre la variable dependiente y las variables independientes, lo que se suele llamar *modelo*. Los modelos de regresión lineal son algunos de los más conocidos y utilizados, dado su simpleza y aplicabilidad. Estos modelos proponen explicar el comportamiento de la variable dependiente a partir de una combinación lineal de las independientes:

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

donde $\beta_1, \beta_2, ..., \beta_p$ son coeficientes desconocidos.

En la práctica se cuenta con datos estructurados en forma de tabla, de tamaño $n \times p$, donde las filas corresponden a n mediciones empíricas de todas las variables involucradas, que llamamos *observaciones*. Así, resulta un sistema de n ecuaciones, para cada $i = 1, ..., n$ es:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{1}$$

y escribiendo el sistema en forma matricial:

$$y = X\beta$$

donde los tamaños son: $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ y $\beta \in \mathbb{R}^p$.

Noten que si tuviésemos que $n = p$, estaríamos en el caso de un sistema lineal dado por una matríz cuadrada, que si fuese invertible, ya aprendieron cómo resolver. Sin embargo, en muchas aplicaciones se trabaja con n mucho mayor a p, por lo que el sistema resulta sobre-determinado y podría no tener solución. Veremos que esto no será un problema.

Dado que si $n > p$ el sistema podría no tener solución exacta, el problema que debemos abordar es el de encontrar una solución que mejor aproxime las ecuaciones del sistema, en algún sentido. Para una elección del vector β, nuestra aproximación de la variable dependiente y, según el modelo, será el producto $X\beta$. El objetivo es entonces elegir β tal que la diferencia entre el vector $X\beta$ y el vector y sea lo más pequeña posible.

Recordamos la *distancia* entre dos vectores u e v de $\mathbb{R}^n$, se denota por $\|u - v\|$ y está dada por la fórmula

$$\|u - v\| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}$$

y también recordamos el *producto escalar* (denominado con el símbolo " $\cdot$ ") entre dos vectores u, v en $\mathbb{R}^n$ está dado por la fórmula

$$u \cdot v = \sum_{i=1}^{n} u_i v_i$$

A partir de la idea de distancia entre vectores como noción de proximidad en la solución, debemos elegir β tal que $\|y - X\beta\|$ sea mínima. A esta solución óptima la llamaremos β^*:

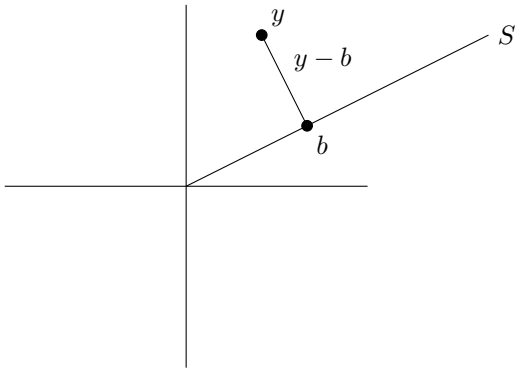
$$\|y - X\beta^*\| = \min_{\beta \text{ en } \mathbb{R}^p} \|y - X\beta\|$$

Por último, introducimos el siguiente resultado central para llegar a una fórmula para β^*:

Teorema. Sea y un vector cualquiera de $\mathbb{R}^n$ y S un subespacio de $\mathbb{R}^n$. El vector de S que minimiza la distancia del subespacio S al vector y es aquel b de S tal que $y - b$ es ortogonal a todo vector s de S. Es decir, las siguientes dos condiciones son equivalentes:

1. $\|y - b\| = \min_{s \in S} \|y - s\|$
2. $(y - b) \cdot s = 0$ para todo s en S

Al vector b se lo denomina la *proyección ortogonal* de y sobre S.



Visualización del teorema en el plano $\mathbb{R}^2$. El segmento $y - b$ es ortogonal a S, es decir, es ortogonal a cualquier vector de S, y es el de menor longitud de y a S.

2. Ejercicios

Primera parte. El objetivo de esta sección es deducir una fórmula para la solución óptima β^* siguiendo los pasos a continuación:

a) Mostrar que el espacio columna de la matriz X es un subespacio vectorial de $\mathbb{R}^n$:

$$Col(X) = \{b \text{ en } \mathbb{R}^n \text{ tales que } b = X\beta \text{ con } \beta \text{ variando en } \mathbb{R}^p\}$$

b) Supongamos que cuando hablamos de vectores en $\mathbb{R}^n$ nos referimos a vectores columna de $\mathbb{R}^{n\times 1}$. Mostrar en ese caso que el producto escalar entre dos vectores u, v en $\mathbb{R}^n$ puede calcularse como:

$$u \cdot v = v^T u$$

donde operación en el lado derecho de la igualdad es el producto de matrices usual.

c) Aplicando el teorema tomando como subespacio S el subespacio del ítem (a), el punto y de $\mathbb{R}^n$ como el vector de la variable dependiente, y el vector b como $b = X\beta^*$, convertir esta ecuación de optimalidad

$$\|y - X\beta^*\| = \min_{\beta \text{ en } \mathbb{R}^p} \|y - X\beta\|$$

en la condición de ortogonalidad que corresponde a la equivalencia 2 del teorema.

d) A la ecuación obtenida en el ítem (c), aplicarle la identidad del producto escalar vista en el item (b), para llegar a la ecuación:

$$X^T(y - X\beta^*) \cdot \beta = \mathbf{0}$$

e) Se sabe que el único vector que es ortogonal a todo vector v de $\mathbb{R}^n$ es el vector nulo. Es decir, si u es un vector fijo tal que $u \cdot v = \mathbf{0}$ para todo v en $\mathbb{R}^n$, entonces $u = \mathbf{0}$. Usando esto y la ecuación obtenida en el ítem (d), llegar a la fórmula:

$$X^T X\beta^* = X^T y$$

f) Finalmente, suponiendo que las columnas de X son linealmente independientes, se tiene que la matriz $X^T X$ es invertible. Despejar β^* de la ecuación del ítem (e) para llegar a la fórmula de la solución óptima al problema de regresión.

Segunda parte. En esta sección la idea es realizar regresión lineal en $\mathbb{R}^2$ y analizar como se comportan las soluciones obtenidas.

1. Usando los datos del archivo *ejercicio_1.csv*:

   a) Graficar todos los puntos en el plano xy.

   Nota: La primer columna del archivo marca el valor de x y la segunda el valor de y de cada punto. Recomendamos usar la biblioteca *pandas* para leer los archivos con la función *read_csv*.

   b) Utilizando los conceptos teóricos desarrollados en la primera parte, hallar la recta que mejor aproxima a los datos.

   c) Realizar nuevamente los incisos (a) y (b) pero considerando los puntos

   $$\{(x_i, y_i + 12) \text{ con } i = 1...n\}$$

   donde (x_i, y_i) eran los puntos originales. ¿Es buena la aproximación realizada?, ¿cuál es el problema?

   d) ¿Cómo se podría **extender** el modelo para poder aproximar cualquier recta en el plano?

2. Usando los datos del archivo *ejercicio_2.csv*:

   a) Graficar y aproximar los puntos con una recta.

b) Imaginemos que los datos forman parte de mediciones de algún tipo, como por ejemplo la temperatura de un procesador a lo largo del tiempo), y queremos predecir cuál va a ser la temperatura en el futuro. ¿Es buena la aproximación que realizamos?, ¿cuál fue el problema en este caso?

Tercera parte. Regresión lineal en datos reales.

En esta sección utilizaremos el conjunto de datos provisto en Machine Learning Repository. Este consiste en datos de ventas de 414 casas en Taiwan. La información provista por casa es (en orden):

i) La fecha en que se realizó la transacción. Expresada en formato

$$\text{año} + \frac{\text{numero\_mes}}{12}$$

ii) La edad de la casa en años.
iii) La distancia a la estación de tren o subte más cercana en metros.
iv) La cantidad de almacenes alcanzables a pie.
v) La latitud en grados.
vi) La longitud en grados.
vii) El precio por Ping. La cual es una unidad utilizada en Taiwan que representa 3,3 metros cuadrados.

Vamos a dividir este conjunto de datos en dos:

i) *Datos de entrenamiento*: usamos los datos desde la observación 1 a la 315 inclusive.
ii) *Datos de test*: usamos los datos desde la observación 316 a la 414 inclusive.

1. Teniendo en cuenta la teoría desarrollada en la primer parte del trabajo práctico y usando los datos de entrenamiento:

   a) Estimar los parámetros $\hat{\beta}$ que minimizan el error cuadrático medio para este problema.

   b) Encontrar $\hat{\mathbf{y}}$ la estimación de la variable de respuesta.

   c) ¿Cuánto vale el error cuadrático medio?

   Definimos error cuadrático medio como

   $$ECM(\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

   donde y_i son observaciones de una variable y $\hat{y_i}$ estimaciones de las mismas.

2. Utilizando los datos de test, analizar cuál es el error cuadrático medio al utilizar los parámetros $\hat{\beta}$ estimados en el punto anterior.

   a) ¿Es la estimación mejor que sobre los datos originales?, ¿a qué se debe la discrepancia?

   b) ¿Qué sucede con el ECM del segundo conjunto de casas si se realiza la regresión sobre todos los datos al mismo tiempo (es decir, las 414 casas)?

3. Graficar el error cometido por cada casa. Es decir el valor absoluto de la diferencia entre el precio por Ping real y el estimado.

4. Imaginemos que se agrega una nueva columna a los datos que informa el año en que la misma fue construida. ¿Disminuiría esto el ECM?

2.1. Condiciones de entrega

Fecha final de entrega: 21 de abril

Modalidad de entrega: Via campus

El trabajo debe realizarse en grupos de a dos. El método de entrega es a través del campus y debe consistir de tres archivos:

1. *informe.pdf*: un informe en el cual se explique el problema, se detalle las tareas realizadas (y cómo fueron resueltas) y los resultados de sus experimentos, es decir graficos, tablas y cualquier otra información que quieran proveer;

2. *codigo.ipynb*: el código *Python* utilizado en formato jupyter-notebook (.ipynb);

3. *codigo.pdf*: la notebook del punto anterior ya corrida y exportada a formato pdf;

Pueden utilizar todas las bibliotecas que usamos en la materia (*numpy*, *matplotlib*, *pandas*), excepto aquellas funciones especificas de regresión lineal o mínimos cuadrados (por ejemplo lstsq).

Las consignas sirven como guía del trabajo, pero esperamos que profundicen más alla de lo que se pide. Expliquen con detalle todas las operaciones matriciales y vectoriales que utilizaron en la primera parte.