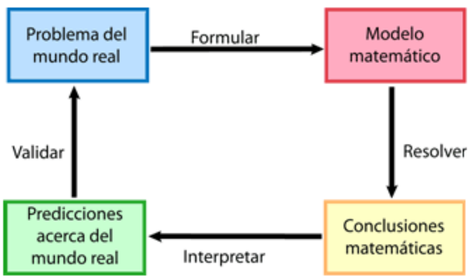
Clase N°7. Modelo de regresión lineal simple.

Un ***modelo*** es útil para representar un fenómeno de la naturaleza. Modelar matemáticamente un problema permite, entre otras cosas, poder comprender el grado en que se relacionan las variables objeto de estudio, predecir su comportamiento ante variaciones imprevistas de su entorno y resolver problemas bajo una mirada cuantitativa.

Desde una perspectiva genérica, un modelo puede ser interpretado como la representación de un suceso, cuya simbolización puede adquirir diferentes formatos: gráficos, esquemas, fórmulas y ecuaciones.

En el caso en el que un modelo pueda ser representado a través de ecuaciones y fórmulas matemáticas se estará haciendo referencia a un *modelo matemático,* donde el pensamiento analítico juega un rol importante ya que permite interpretar una situación particular de la vida cotidiana.

A continuación se muestra un esquema representativo con las diferentes etapas de un modelo matemático.



Según la información en la que se basa el modelo es posible discriminar *entre modelo heurístico,* en el que la información tiene sostén en las definiciones teóricas las de las causas o motivos naturales que generan el fenómeno que se pretende estudiar y *modelo empírico* basado en el estudio de resultados experimentales.

Por otra parte, basados en el tipo de resultado que se pretende obtener, es posible discernir entre *modelos cualitativos* y *modelos cuantitativos.* Los modelos cualitativos no buscan un resultado exacto sino una tendencia, como por ejemplo determinar si cierto parámetro del problema se incrementa o disminuye, este tipo de modelos suele valerse de información gráfica. En cambio los modelos cuantitativos buscan como resultado valores numéricos lo más preciso posible.

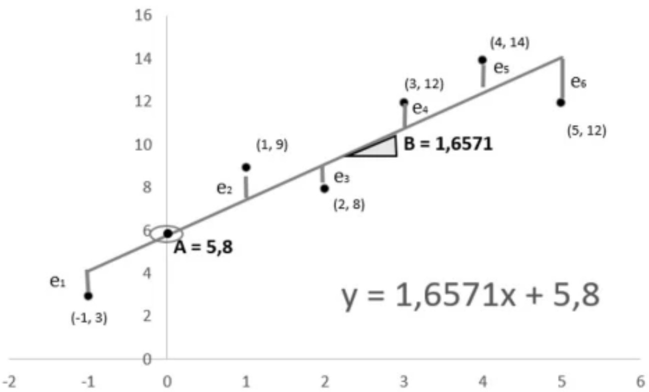
Modelo de regresión lineal simple

La ***regresión lineal*** es un modelo matemático utilizado para aproximar la relación de dependencia entre dos variables, la variable independiente o *variable observable* que es la que surge de la observación de un proceso a estudiar y la variable dependiente o *variable respuesta* que expresa los valores que se infieren a través de la variable observable.

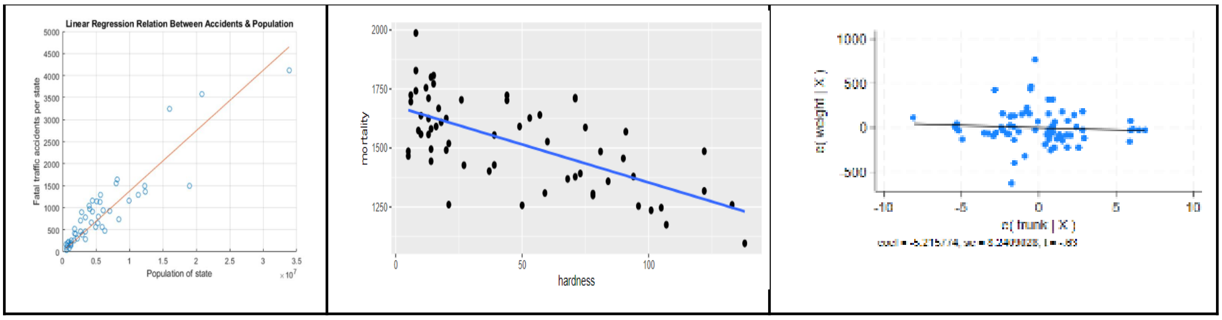
Dado un conjunto de datos representados sobre una nube de puntos, se busca la recta que mejor los modela, denominada ***recta de regresión lineal.*** Esta permite describir cuantitativamente el grado de asociación, o la correlación existente, entre las dos variables involucradas en el fenómeno estudiado. La recta se representa mediante la sig. ecuación:

$$y = \beta_o + \beta_1 . x + \epsilon,$$

donde los parámetros β_o , β_1, ϵ representan la ordenada al origen, la pendiente de la recta y el error cometido en la aproximación, como se muestra en el sig. ejemplo:

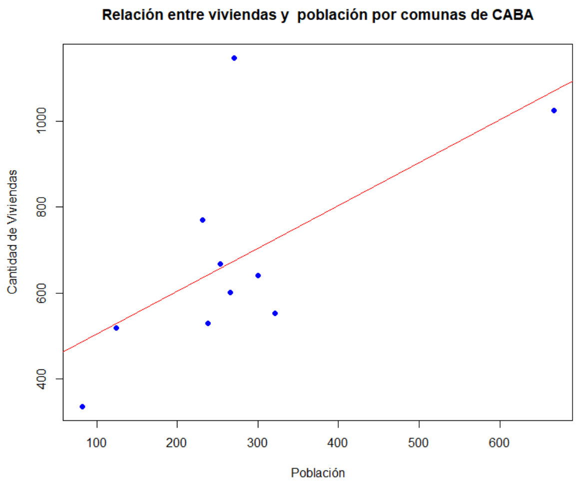


Dependiendo el valor que tome la pendiente de la recta, la recta será creciente si la pendiente es positiva, decreciente si la inclinación de la recta es negativa y constante en caso que la pendiente sea nula.



Ejemplo 1.

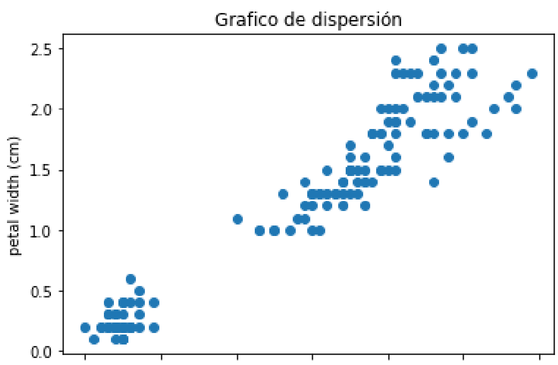
Se quiere analizar la relación entre la cantidad de viviendas y la cantidad de población en una comuna de la ciudad de Buenos Aires.



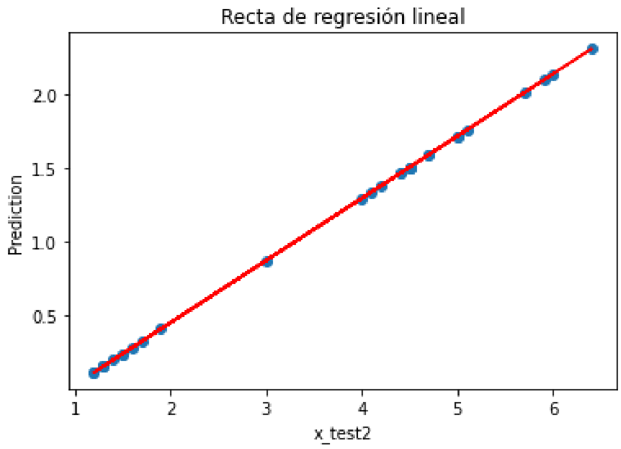
Si bien los datos presentan un alto grado de dispersión, la recta de regresión lineal permite estimar una relación lineal positiva pero débil entre las variables analizadas.

Ejemplo 2.

En base al set de datos iris se quiere predecir el ancho del pétalo en relación al largo del pétalo. En primer lugar se representa el grado de dispersión de los datos entre las variables ancho y largo del pétalo:



En el grafico se puede notar que la nube de dispersión muestra una tendencia ascendente. Luego, al implementar un modelo de regresión lineal se estima la recta que mejor aproxima ambas variables, como se muestra a continuación:



La recta de regresión lineal presenta una pendiente positiva, lo cual muestra una relación directa y fuerte entre las variables analizadas, es decir, al aumentar el tamaño del largo del pétalo aumenta la dimensión del ancho del pétalo.

Hiperparámetros del modelo en Python:

Con la opción *modelo.get.param* () se pueden obtener los hiperparámetros del modelo de regresión lineal.

`'copy_X': True.` Copia el resultado en pantalla.
`'fit_intercept': True.` Calcula el coeficiente para el modelo.
`'n_jobs': None.` Estable el número de jobs para el modelo.
`'normalize': 'deprecated'.`No está disponible la función para normalizar la curva.
`'positive': False.` Fuerza a que el coeficiente del modelo tome un valor positivo.

Métricas en Python

Al igual que en los árboles de regresión se pueden utilizar las siguientes métricas

MAE: The Mean Absolute Error.

MSE: The Mean Squared Error.

RMSE: The Root Mean Squared Error.

Adicionalmente se estima el coeficiente de determinación para medir el grado de correlación entre las variables.

El coeficiente de determinación (R^2).

En el caso de un modelo lineal, el valor del R^2 es un estadístico que mide la bondad de ajuste del modelo y el cual se obtiene a partir del cuadrado del coeficiente de correlación lineal de Pearson. Muestra que porcentaje de la variable respuesta es explicada por el modelo. El coeficiente de determinación varía entre 0 y 1. Los valores próximos a uno reflejan un ajuste lineal perfecto, es decir, la correlación entre las variables es fuerte.

Actividad: Preguntas de opción múltiple.

Responder con verdadero o falso a los siguientes enunciados:

1. Un modelo es utilizado para representar un fenómeno de la naturaleza.
2. Un modelo de regresión lineal simple analiza la relación entre dos variables observables y una variable respuesta.
3. El error mide la dispersión entre el valor real y el valor predicho por el modelo.