
Correlaciones y Regresiones

Correlaciones y Regresiones

Correlaciones

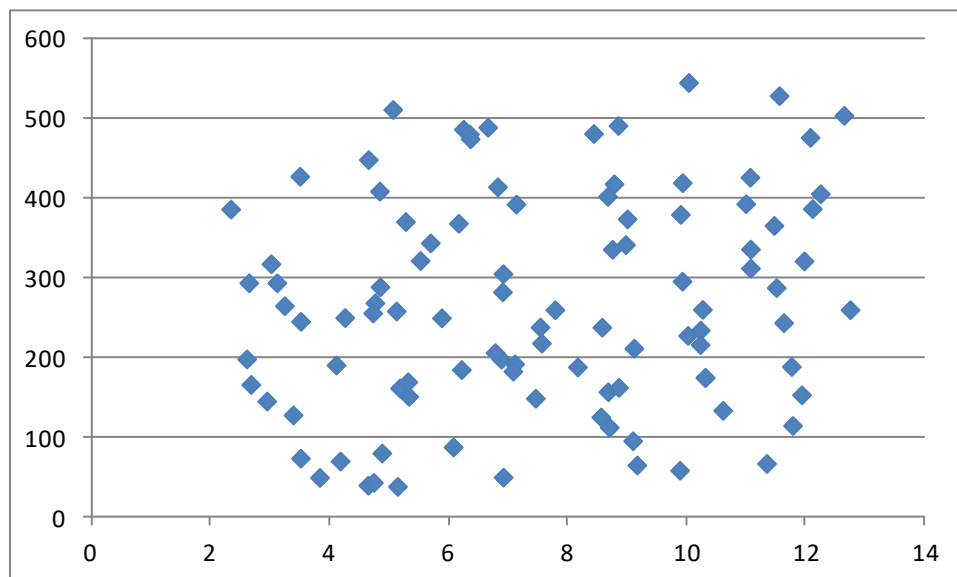
Muchas veces nos encontramos que dos variables aleatorias parecen tener un comportamiento asociado. Por ejemplo, cuando una sube la otra también lo hace. A ese comportamiento lo llamamos **correlación**.

No debemos confundir correlación con causalidad. De hecho las estadísticas no tienen, de por sí, nada para decir sobre la causalidad. Si tenemos dos variables aleatorias correlacionadas A y B tanto podría A ser causa de B, como B causa de A como ambas consecuencias de una tercera variable C.

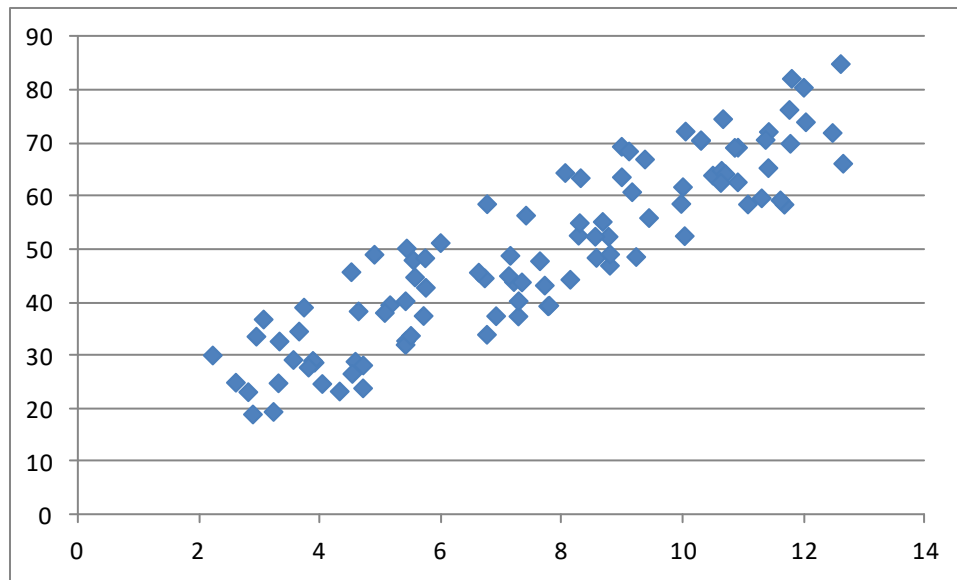
Vamos a descubrir más adelante que en una situación real las variables involucradas se multiplican elevando rápidamente la complejidad de un problema. Una forma de controlar esta situación es reducir la cantidad de variables dejando sólo las que no están correlacionadas.

Dos variables no correlacionadas son habitualmente llamadas independientes entre sí. Tomemos el caso de 100 muestras de dos variables aleatorias X e Y, que se obtienen simultáneamente 100 veces:

Caso A:



Caso B:



En el caso A no parece haber correlación alguna entre los valores que va tomando X y los que va tomando Y.

En el caso B parece haber cierta tendencia a que los valores de X más grandes estén asociados a los valores más grandes de Y.

Para obtener una variable representativa de la correlación se calcula:

$$C = \frac{\sum_{i=1}^{100} (x_i - x_{medio}) * (y_i - y_{medio})}{\sqrt{\sum_{i=1}^{100} (x_i - x_{medio})^2} * \sqrt{\sum_{i=1}^{100} (y_i - y_{medio})^2}}$$

Para el caso A da 0.13 mientras que para el caso B da 0,91

Un valor próximo a 1 habla de una correlación positiva, un valor cercano a -1 habla de una correlación negativa. Un valor cercano a cero nos indica que no podemos estar seguros de que exista correlación alguna.

Ejercicio 4.1

Busque por internet los 10 países con mayor PBI y, para cada uno de ellos su PBI y su población.

¿Existe alguna correlación entre el PBI y la población?

Luego calcule el PBI per cápita.

¿Existe alguna correlación entre el PBI per cápita y la población?

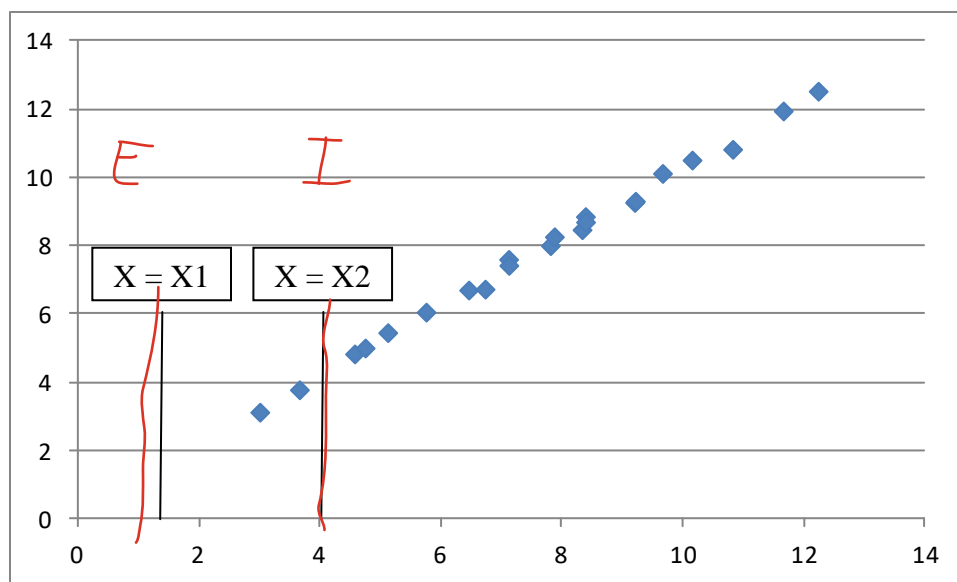
Regresiones

Muchas veces nos encontramos con fenómenos donde dos variables parecen correlacionadas. En ese caso nos interesaría poder calcular cual sería el valor de una variable para un valor dado de la otra.

A la variable cuyo valor vamos a predecir la llamaremos variable dependiente y la notaremos Y . A la otra variable la llamaremos independiente y la notaremos X . Esta identificación es, en principio, arbitraria. Ya discutiremos más adelante las implicancias de elegir como independiente una variable u otra.

Interpolaciones y extrapolaciones:

Hay dos modalidades muy distintas de este problema. La modalidad más fácil corresponde a predecir el valor de la variable dependiente para un valor de la variable independiente que está rodeado por los datos que tenemos:



Para el caso de $X = X2$ tenemos un ejemplo de lo que llamamos interpolación ya que hay valores de X mayores y menores a $X2$.

En cambio, para el caso de $X = X1$, no hay valores de X para los cuales tenga datos que sean menores a $X1$, sólo mayores, por lo tanto se trata de una extrapolación.

Es muy importante tener claro que una interpolación suele ser mucho más confiable que una extrapolación.

Regresión Lineal:

En un caso como el de la figura anterior parece bastante claro que los datos se concentran alrededor de una recta.

La regresión lineal consiste en determinar cuál es la recta que mejor representa al conjunto de puntos.

¿Qué quiere decir, en términos matemáticos, representar mejor al conjunto de puntos? Típicamente los valores independientes son mejor conocidos que los dependientes. Por ese motivo es habitual tomar a los datos independientes como libres de error y pensar que todo el error se concentra en los datos dependientes.

Veamos los elementos del problema:

Tenemos un conjunto de N puntos:

$$\{(X_i; Y_i) \text{ tal que } 1 \leq i \leq N\}$$

Tenemos la recta candidata:

$$Y = aX + b$$

¿Cuánto se equivoca la recta respecto de cada punto?

$$E_i = Y_i - (aX_i + b)$$

Este número será a veces positivo y a veces negativo. Para que no se compensen entre ellos voy a sumar sus cuadrados:

$$\chi^2 = \sum_{i=1}^N (Y_i - aX_i - b)^2$$

Variando a y b puedo tener la distancia de mi conjunto de puntos a todas las rectas posibles. ¿Cuál de todas las rectas posibles me interesa? La que hace mínima la distancia.

Para los que hayan visto algo de cálculo lo que debo hacer es resolver el siguiente sistema:

$$\begin{cases} \frac{\partial \chi^2}{\partial a} = 0 \\ \frac{\partial \chi^2}{\partial b} = 0 \end{cases}$$

Para expresar convenientemente la solución conviene definir los siguientes objetos intermedios donde N es el número de puntos del que disponemos en nuestra muestra:

$$\sigma_{jl} = \sum_{i=1}^N X_i^j Y_i^l$$

Con eso obtenemos:

$$a = \frac{\sigma_{10}\sigma_{01} - N\sigma_{11}}{\sigma_{10}^2 - N\sigma_{20}}$$

$$b = \frac{\sigma_{01} - a\sigma_{10}}{N}$$

Además podemos estimar las incertidumbres con las que conocemos a y b:

$$\Delta a^2 = \frac{N}{N\sigma_{20} - \sigma_{10}^2} \frac{\chi^2(a, b)}{N - 2}$$

$$\Delta b^2 = \frac{\sigma_{20}}{N\sigma_{20} - \sigma_{10}^2} \frac{\chi^2(a, b)}{N - 2}$$

Además podemos definir el coeficiente de correlación:

$$R^2 = \frac{(N\sigma_{11} - \sigma_{10}\sigma_{01})^2}{(N\sigma_{20} - \sigma_{10}^2)(N\sigma_{02} - \sigma_{01}^2)}$$

Tal y como lo hemos definido R^2 tiende a 1 cuando estamos trabajando con una recta bien definida y se aproxima a cero cuando tenemos una nube de puntos que difícilmente pueda aproximarse por una recta.

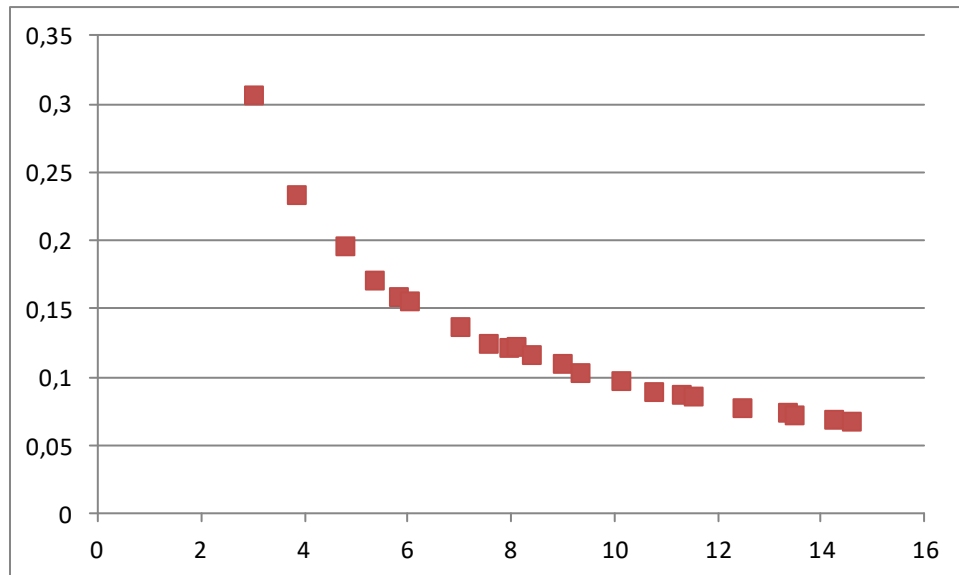
Cuando $|a| < \Delta a$ nos da la pauta que la recta en cuestión sería compatible con una horizontal.

Cuando $|b| < \Delta b$ nos indica que la recta en cuestión es compatible con pasar por el punto (0,0)

Algo que a veces resulta muy importante es cuan cerca de 1 queda R^2 . Se debe tener cuidado pues un valor de R^2 cercano a 0.9 puede ser muy bueno para correlaciones en circunstancias comerciales y muy malo para un trabajo experimental en mecánica elemental.

Linealización de regresiones:

A veces los puntos no se presentan agrupados de manera que se parezcan a una recta aunque, un ojo entrenado puede ver una forma funcional que pueda identificar:



Se puede atacar este problema con herramientas más complejas que veremos en el último punto de esta unidad. Sin embargo por ahora, vamos a proponer hacer un cambio de variable del tipo:

$$Z_i = \frac{1}{Y_i}$$

Entonces, calculamos los nuevos valores de la variable dependiente (Ahora Z_i) y aplicamos lo que ya sabemos de regresión lineal.

Este truco sirve para extender lo que ya sabemos de regresión lineal a muchas circunstancias comunes sin complicarnos con técnicas más generales.

Regresión cuadrática

Veamos los elementos del problema:

Tenemos un conjunto de N puntos:

$$\{(X_i; Y_i) \text{ tal que } 1 \leq i \leq N\}$$

Tenemos la recta candidata:

$$Y = aX^2 + bX + c$$

Cuanto se equivoca la recta respecto de cada punto:

$$E_i = Y_i - (aX_i^2 + bX_i + c)$$

Este número será a veces positivo y a veces negativo. Para que no se compensen entre ellos voy a sumar sus cuadrados:

$$\chi^2 = \sum_{i=1}^N (Y_i - (aX_i^2 + bX_i + c))^2$$

Variando a, b y c puedo tener la distancia de mi conjunto de puntos a todas las rectas posibles. ¿Cuál de todas las rectas posibles me interesa? La que hace mínima la distancia.

Para los que hayan visto algo de cálculo lo que debo hacer es resolver el siguiente sistema:

$$\begin{cases} \frac{\partial \chi^2}{\partial a} = 0 \\ \frac{\partial \chi^2}{\partial b} = 0 \\ \frac{\partial \chi^2}{\partial c} = 0 \end{cases}$$

Recordando los objetos intermedios ya definidos:

$$\sigma_{jl} = \sum_{i=1}^N X_i^j Y_i^l$$

Se trata sólo de resolver el sistema:

$$\begin{pmatrix} \sigma_{00} & \sigma_{10} & \sigma_{20} \\ \sigma_{10} & \sigma_{20} & \sigma_{30} \\ \sigma_{20} & \sigma_{30} & \sigma_{40} \end{pmatrix} x \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sigma_{01} \\ \sigma_{11} \\ \sigma_{21} \end{pmatrix}$$

El cálculo de a, b y c se deja como ejercicio para el estudiante entusiasta.

Regresión polinómica:

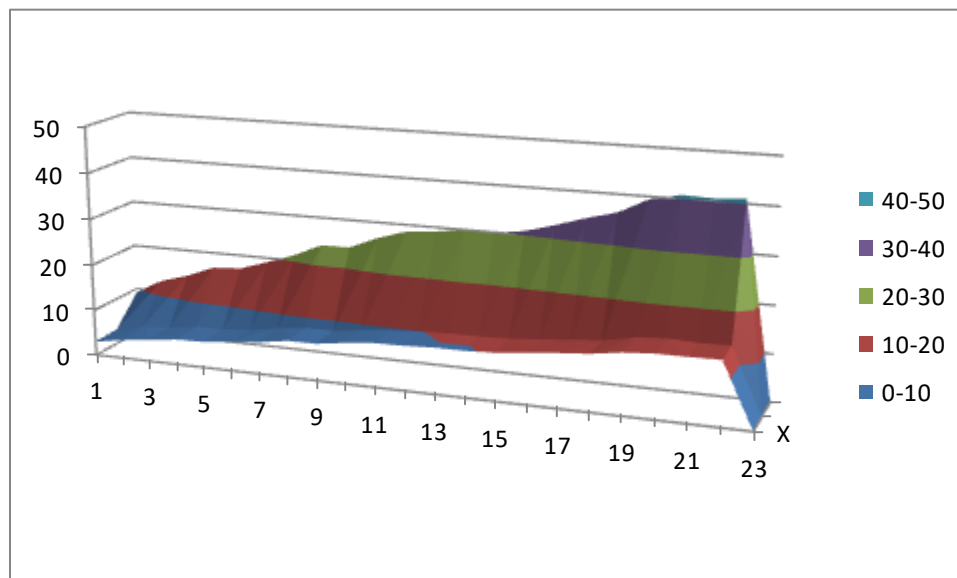
Cómo resultará obvio a estas alturas lo hecho para la regresión cuadrática puede extenderse sin problemas a polinomios de grado superior.

Es necesario, sin embargo, realizar una advertencia. No conviene aumentar el grado sin tener noticia que cada término tiene un significado inmediato y propio del problema que se trata de resolver. Caso contrario se corre el riesgo de que se produzca un "sobreajuste" (overfitting en inglés) que disminuya incluso el valor de las interpolaciones.

Las funciones polinómicas de alto grado pueden oscilar rápidamente para acomodarse a detalles caprichosos de los datos. Estas oscilaciones pueden hacer que los valores que resulten incluso de las interpolaciones carezcan de poder predictivo para el problema concreto sobre todo si el grado asignado al polinomio no corresponde a alguna forma de dependencia o ley que los datos debieran, por algún motivo, seguir.

Regresión bilineal

A veces nos encontramos con un conjunto de ternas donde el último dato parecería ser función de los otros dos. Este tipo de situaciones debería reflejarse en un gráfico en tres dimensiones:



Veamos los elementos del problema:

Tenemos un conjunto de N puntos:

$$\{(X_i; Y_i; Z_i) \text{ tal que } 1 \leq i \leq N\}$$

Tenemos el plano candidato:

$$Z = aX + bY + c$$

Cuanto se equivoca el plano respecto de cada punto:

$$E_i = Z_i - (aX_i + bY_i + c)$$

Este número será a veces positivo y a veces negativo. Para que no se compensen entre ellos voy a sumar sus cuadrados:

$$\chi^2 = \sum_{i=1}^N (Z_i - (aX_i + bY_i + c))^2$$

Variando a, b y c puedo tener la distancia de mi conjunto de puntos a todas las rectas posibles. ¿Cuál de todas las rectas posibles me interesa? La que hace mínima la distancia. Aplicando el truco ya conocido por el lector a estas alturas se pueden despejar a, b y c:

$$\left\{ \begin{array}{l} \frac{\partial \chi^2}{\partial a} = 0 \\ \frac{\partial \chi^2}{\partial b} = 0 \\ \frac{\partial \chi^2}{\partial c} = 0 \end{array} \right\}$$

Esto resulta en un sistema de 3 ecuaciones con tres incógnitas. Su solución se deja, de nuevo, como ejercicio para el lector.

Esto puede extenderse a regresiones multilineales sin grandes dificultades.

Regresión general

Cuando la forma funcional es compleja existen todavía procesos para realizar el ajuste. Sin detenernos a describir con precisión su forma matemática vamos a realizar una discusión somera de las ventajas y desventajas relativas.

El primer método que vamos a mencionar es el plasmado en el algoritmo "Curve Fit" Este algoritmo, para funcionar eficientemente requiere que el cálculo del valor de la función en un punto sea rápido ya que para cada dimensión calcula la derivada parcial en forma numérica para la función χ^2 y trata de orientarse hacia el mínimo por un camino de aproximaciones sucesivas.

Si la función tiende a variar muy violentamente será preciso explorarla siempre con pasos que resulten menores a las distancias características de variación en cada dimensión para asegurar la convergencia.

Esto puede conspirar contra la rapidez del hallazgo de la solución. Hay que notar que por cada solución candidata (o paso de la optimización) se calculan 2 M valores donde M es la cantidad de parámetros de la función a ajustar.

Si el cálculo de un punto es largo entonces se puede recurrir al método de la ameba. Este parte de M+1 valores (donde M es el número de parámetros a obtener por medio de la optimización del ajuste)

Cada uno de los valores corresponde a los vértices de un hiper-tetraedro que vive en el espacio de los parámetros a optimizar. Se elijen el mínimo y el máximo de los $M+1$ valores y se trata de cambiar el valor del máximo alargando, acortando o reflejando el punto del máximo respecto de la cara formada por los otros M puntos.

Este método resulta más eficiente cuando el cálculo del valor de la función es lento ya que por cada solución candidata se calculan unos pocos valores.

Medias Móviles:

Como forma de interpolar se utilizan, a veces, las medias móviles. Esto puede hacerse en una dimensión o en dos.

Si la dimensión en la que se hace la media móvil es el tiempo y existe, por ejemplo, un ciclo semanal en la variable que se está promediando la unidad de promedio debe ser un múltiplo entero de la semana. Lo contrario lleva a ver oscilaciones que son un subproducto del ciclo semanal.

Otra advertencia importante para las medias móviles es que presentan la información "retrasada" en medio ciclo. Esto puede fácilmente llevar a problemas de interpretación cuando se trata de encontrar las posibles causas de una variación.

Cálculo de regresiones en R con `lm`

Vamos a aprender a usar la función `lm` que nos provee R para realizar regresiones.

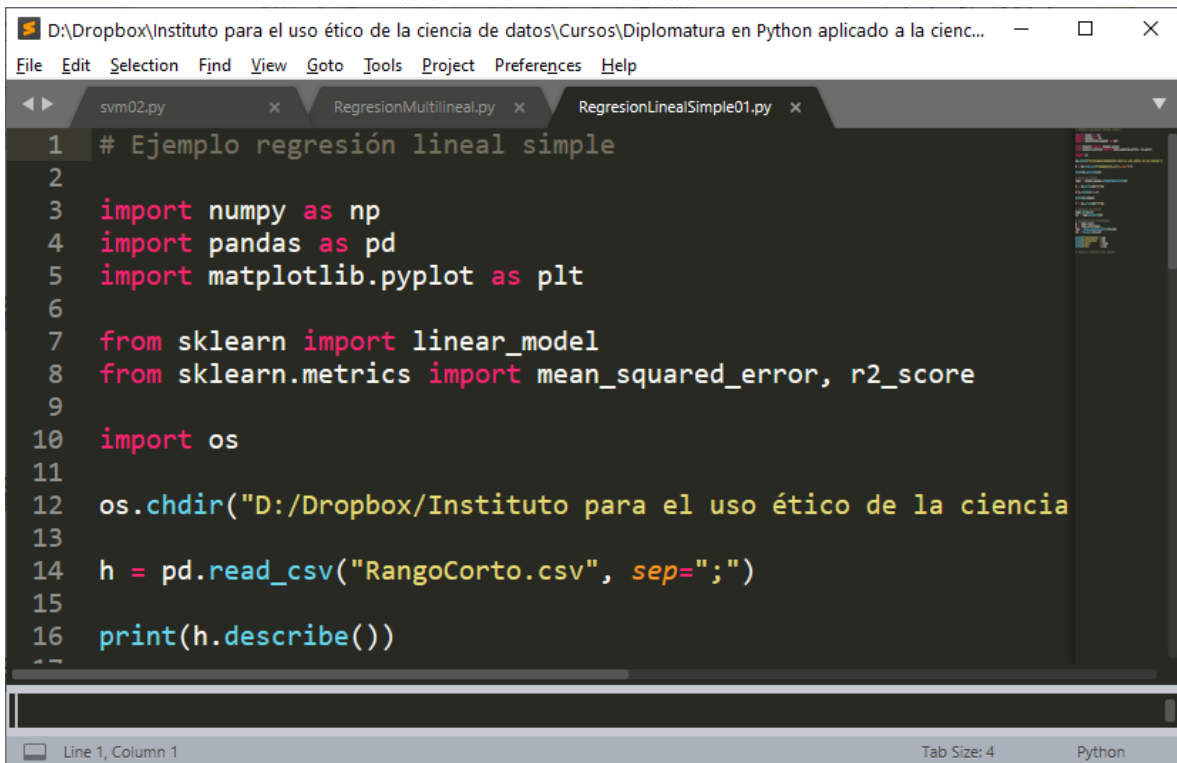
Como toda función vamos a empezar por sus parámetros. En el caso de `lm` esto es fácil, tiene dos parámetros principales. Uno son los datos que va a utilizar y el otro corresponde a la ley que proponemos para vincular las variables que queremos hacer pasar por el proceso de regresión.

Como datos podemos utilizar cualquier dataframe que contenga en las columnas los datos sobre los que vamos a realizar la regresión.

Empecemos por el caso más simple en el cual nos sintamos cómodos. Todo empieza por cargar los datos que vamos a usar:

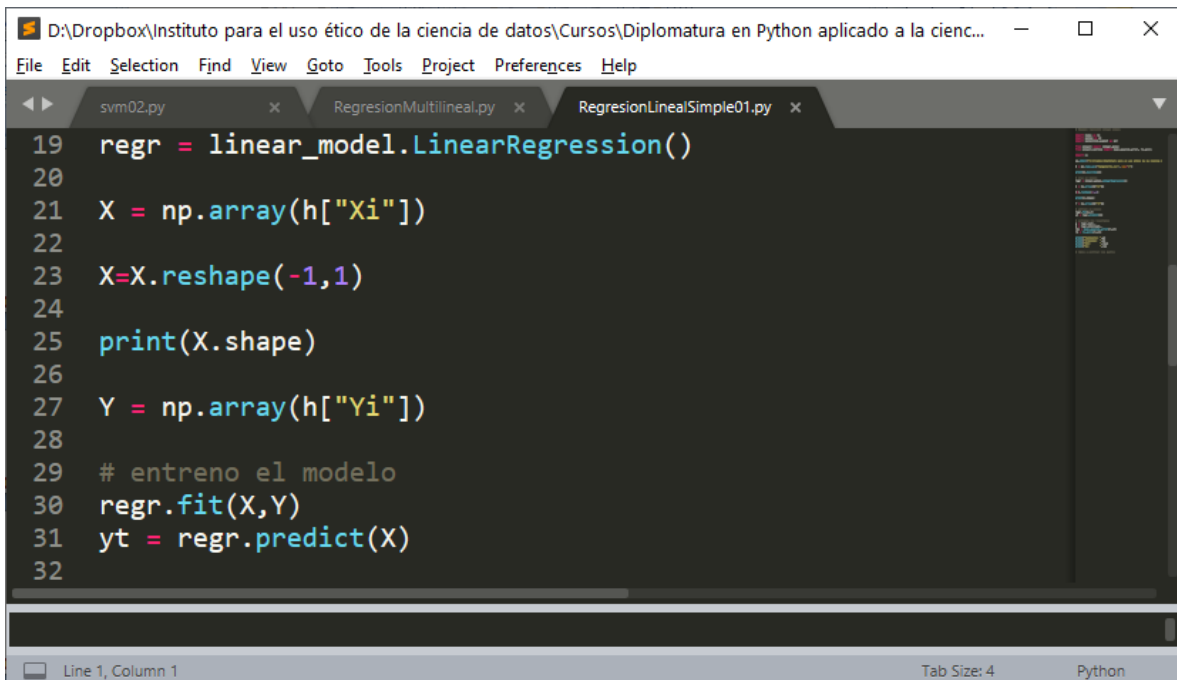
X	Y
1	8,05
2	11,06
3	14,06
4	17,04
5	20,09
6	23,01
7	26,05
8	29,02
9	32,08
10	35,04
11	38,02
12	41,08
13	44,09
14	47,10
15	50,03
16	53,05
17	56,04
18	59,09
19	62,06
20	65,08

Empiezo por importar las librerías que voy a necesitar y cargo el archivo con los datos:



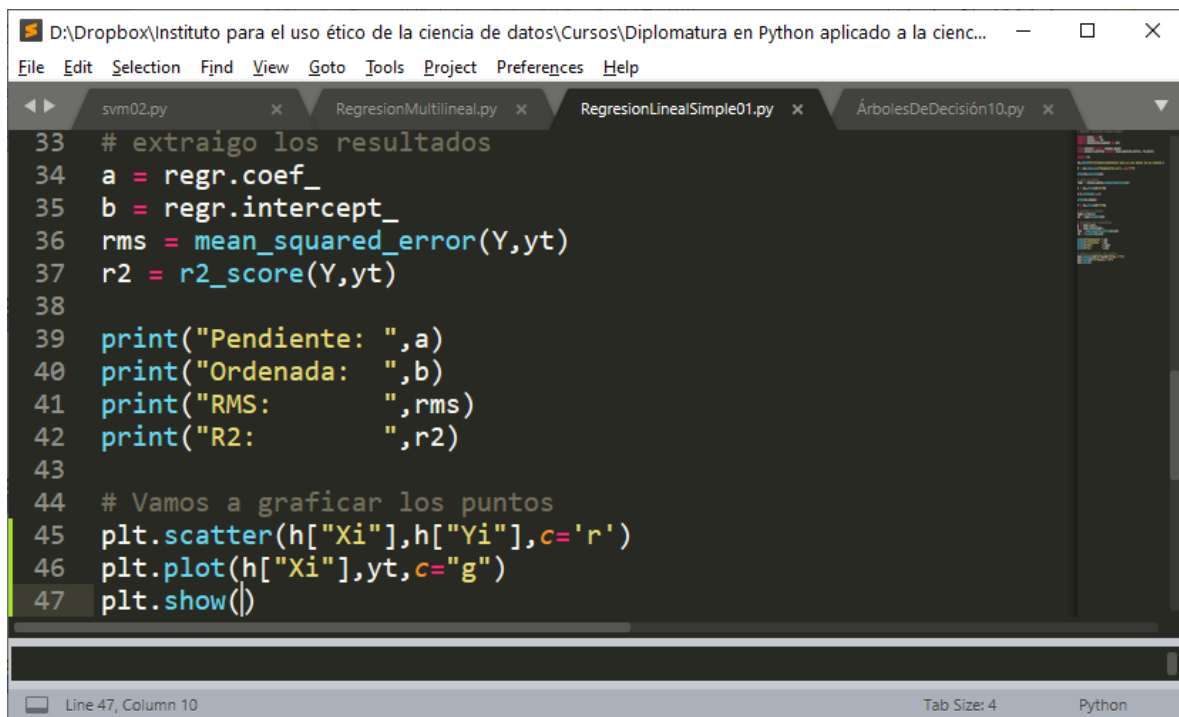
```
1 # Ejemplo regresión lineal simple
2
3 import numpy as np
4 import pandas as pd
5 import matplotlib.pyplot as plt
6
7 from sklearn import linear_model
8 from sklearn.metrics import mean_squared_error, r2_score
9
10 import os
11
12 os.chdir("D:/Dropbox/Instituto para el uso ético de la ciencia
13
14 h = pd.read_csv("RangoCorto.csv", sep=";")
15
16 print(h.describe())
```

Construyo el modelo lineal para lo cual tengo que ajustar la configuración de X que espera dos columnas con datos y así producir la predicción:



```
19 regr = linear_model.LinearRegression()
20
21 X = np.array(h["Xi"])
22
23 X=X.reshape(-1,1)
24
25 print(X.shape)
26
27 Y = np.array(h["Yi"])
28
29 # entreno el modelo
30 regr.fit(X,Y)
31 yt = regr.predict(X)
32
```

Procedemos entonces a extraer los resultados:

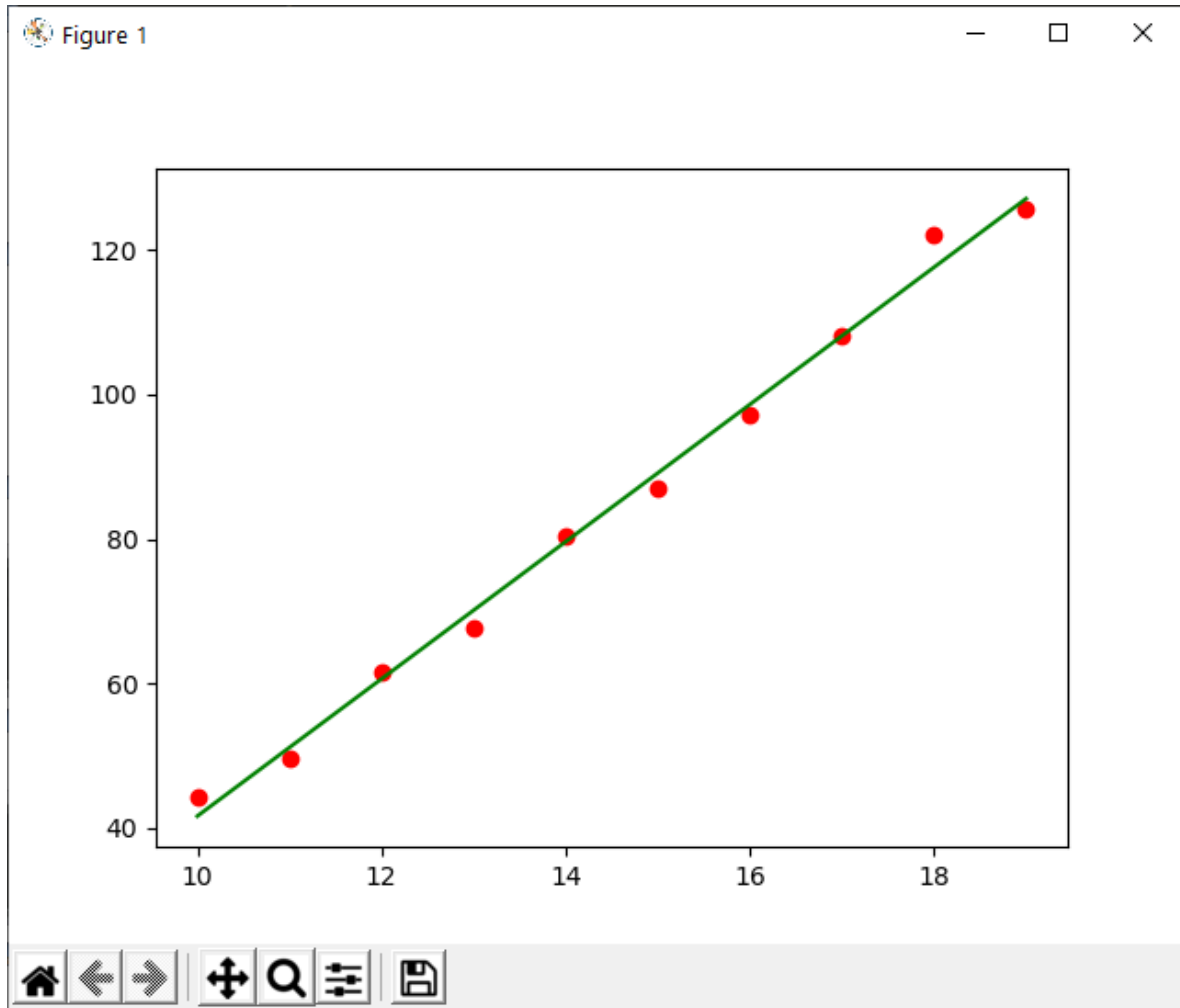


The screenshot shows a Python IDE window with the title bar "D:\Dropbox\Instituto para el uso ético de la ciencia de datos\Cursos\Diplomatura en Python aplicado a la ciencia...". The menu bar includes File, Edit, Selection, Find, View, Goto, Tools, Project, Preferences, and Help. The tab bar shows four open files: svm02.py, RegresionMultilineal.py, RegresionLinealSimple01.py, and ÁrbolesDeDecisión10.py. The main editor displays the following Python code:

```
33 # extraigo los resultados
34 a = regr.coef_
35 b = regr.intercept_
36 rms = mean_squared_error(Y,yt)
37 r2 = r2_score(Y,yt)
38
39 print("Pendiente: ",a)
40 print("Ordenada: ",b)
41 print("RMS:      ",rms)
42 print("R2:       ",r2)
43
44 # Vamos a graficar los puntos
45 plt.scatter(h["Xi"],h["Yi"],c='r')
46 plt.plot(h["Xi"],yt,c="g")
47 plt.show()
```

The status bar at the bottom indicates "Line 47, Column 10", "Tab Size: 4", and "Python".

Para poder graficar los puntos:



En clase también veremos como realizar regresiones multilineales, linealizadas y polinómicas.