*Clase 12. Introducción al análisis predictivo con Regresión*

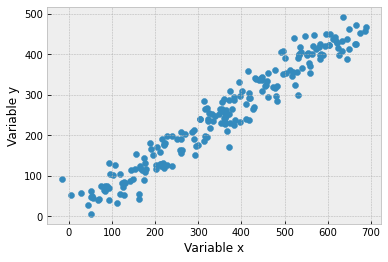
# Acercándonos al concepto de regresión.

## Correlación, causalidad y dependencia.

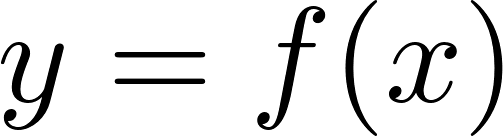
En la clase 7, correspondiente a las herramientas de visualización, estuvimos viendo la posibilidad de que los datos tuvieran algún tipo de relación. Hicimos en ese momento énfasis en la idea de que una relación lineal, que puede medirse por un índice o coeficiente de correlación, no implica necesariamente una causalidad, esto es, que una variable dependa de la otra.

Ahora bien, para esta clase tomemos como punto de partida que de algún modo planteamos la hipótesis de que *podría* existir algún tipo de dependencia de una variable con respecto a la otra. Si este tipo de dependencia existe, queremos ver de qué forma se da esa relación. Tengamos en cuenta que el planteo de la posibilidad de relación es una tarea externa a la formulación del modelo que vamos a ver aquí.

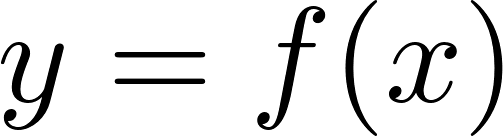
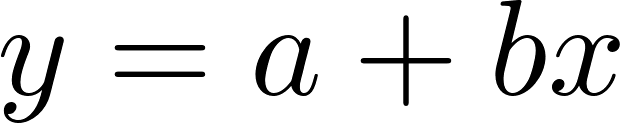
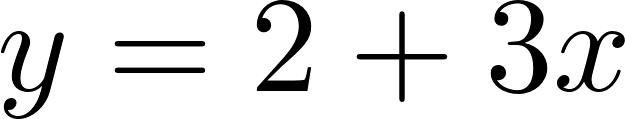
Supongamos entonces que tenemos dos variables, x e y, y veamos el gráfico siguiente.



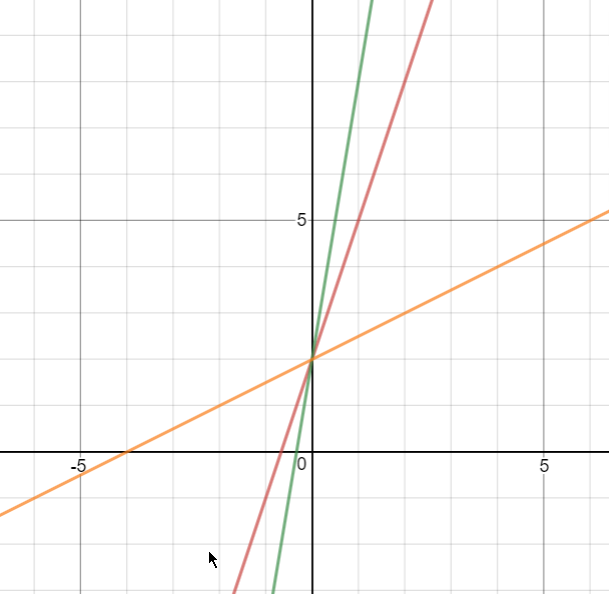
En este caso, pareciera que las variables tienen una fuerte correlación positiva, y si lo pensamos en términos de dependencia, quiere decir que cuando la variable x aumenta, entonces también lo hace la variable y, y viceversa. Notemos aquí la forma diferente de mencionar esta situación, cuando planteamos que ante un cambio en la variable x se produce un cambio en la variable y. A esto lo llamaremos dependencia de la variable y hacia la variable x.

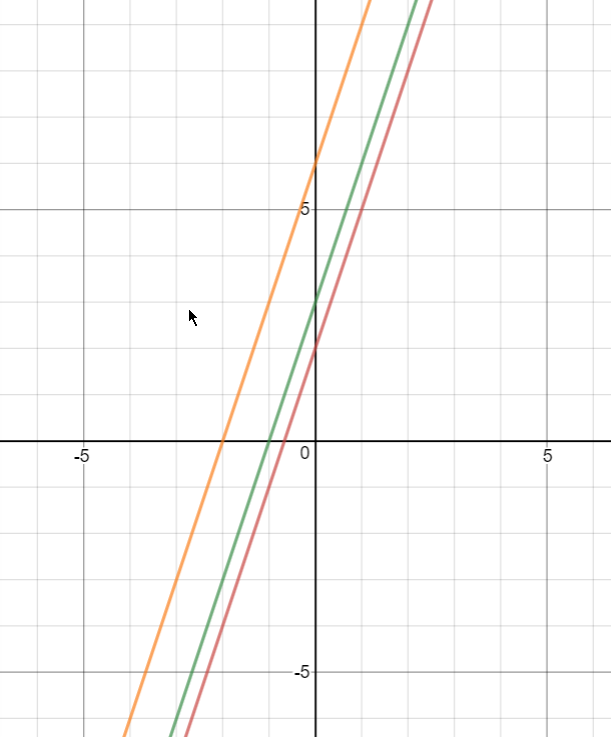
Podemos pensar esta situación como una función matemática estándar [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%20f(x)%20#0), donde la variable y es una función de x, o sea que en definitiva y depende del cambio de x. Otra forma de decir lo mismo es que x es una variable independiente, o sea que su cambio no depende de nuestro modelo, y por su parte los cambios que se den en la variable y se dan de acuerdo a los cambios de x. Aquí entonces estudiaremos la relación entre x e y, teniendo en cuenta cuál es la variación de y ante los cambios de x.

## Función lineal

Repasemos aquí entonces el concepto de función lineal de acuerdo a la función genérica presentada anteriormente [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%20f(x)%20#0). Una función lineal tiene la forma [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%20a%20%2B%20bx%20#0), donde a y b son números reales, por ejemplo [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%202%20%2B%203x%20#0). Esta función genera una recta en el plano. El valor de a (ordenada al origen) muestra cuál es el valor de y cuando x vale 0. En términos gráficos, muestra dónde está “enganchada” la recta en el eje y. El valor de b (pendiente), por su parte, indica el grado de inclinación de la recta. Una recta totalmente horizontal tiene una pendiente igual a cero, mientras que una recta inclinada en el sentido de la correlación positiva tiene una pendiente positiva. Una recta inclinada en el sentido de la correlación negativa tiene una pendiente negativa. Una recta vertical tiene pendiente infinita (aunque no nos preocupemos, porque no va a ser necesario que veamos este último caso).

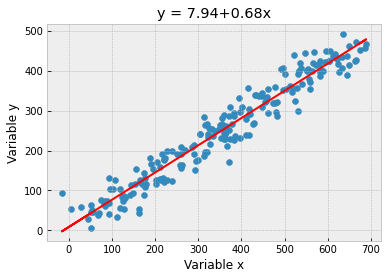
En las figuras siguientes podemos ver, en primer lugar, el mismo valor de a con distintos valores de b, aquí cambia la pendiente o inclinación. En segundo lugar, podemos ver un valor fijo de b para distintos valores de a, aquí cambia la posición de la recta pero su inclinación permanece igual. En cada caso podemos ver entonces qué parte de la función es la que cambia.





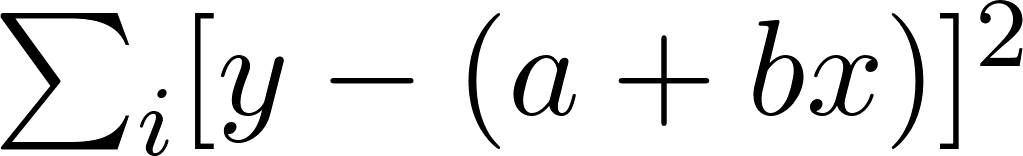
## Ideas previas

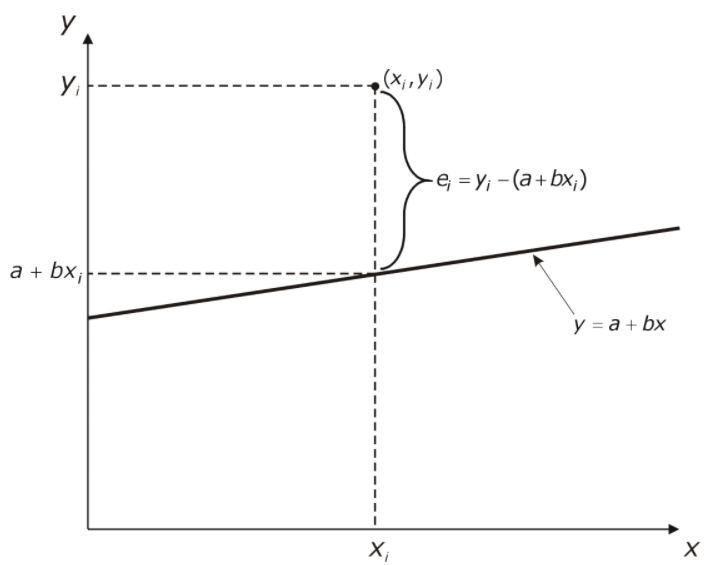
Unamos ahora los dos conceptos. Si tenemos un conjunto de puntos en las variables x e y, y de alguna forma y depende de x, una forma de resumir esta información es suponer que podemos trazar una recta que de alguna manera puede representar a esos puntos. Es claro que no existe una recta que pueda pasar por absolutamente todos los puntos, porque hay determinado grado de dispersión en estos puntos que no permite esto. No obstante, lo que sí podemos hacer es tomar un criterio para la representación, y trazar una recta que cumpla con este criterio. Por ejemplo, una recta que pase “lo más al centro posible” del conjunto de puntos. En este caso, nos encontraremos en el caso de la siguiente figura, donde tenemos una recta que pasa aproximadamente por el centro de la forma dada por los puntos.

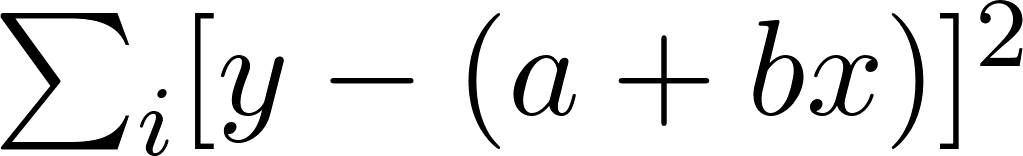


Lo que hacemos aquí es realizar un ajuste de la recta a los datos. A la técnica que utilizamos para realizar este ajuste a un conjunto de puntos por parte de una recta la llamaremos “método de mínimos cuadrados”. Esta recta tiene una forma precisa de construirse que veremos a continuación, y todo conjunto de puntos, cualquiera sea su forma, tiene una recta que se ajuste a ese conjunto por medio del método de mínimos cuadrados. Si este ajuste además cumple con ciertas condiciones, podemos decir que aplicamos un modelo de regresión lineal simple.

Veamos estas ideas. Un punto importante es que un modelo de regresión lineal no es simplemente una recta de ajuste por mínimos cuadrados, sino que deben cumplirse una serie de condiciones rigurosas que deben probarse matemáticamente. Dicho de otra forma, una recta de ajuste por sí sola no conforma un modelo de regresión. A los efectos prácticos del Data Science, veremos las dos condiciones más importantes, denominadas “test de beta” y “coeficiente de determinación”. En la práctica, en la mayoría de los casos, con verificar estas dos condiciones, es válido afirmar que el modelo se comporta de una manera adecuada. Usaremos estas dos condiciones de aquí en adelante.

¿Cómo funciona el método de mínimos cuadrados?. Se toma cada punto individual y se calcula su distancia vertical a la recta (denominada error y simbolizada con la letra e). Se realiza entonces la suma de todas las distancias verticales elevadas al cuadrado. En fórmula [](https://www.codecogs.com/eqnedit.php?latex=%20%5Csum_i%20%5By%20-%20(a%2Bbx)%5D%5E2%20#0). Un ejemplo del error se muestra en la figura.



Matemáticamente, existe una fórmula (que no veremos aquí) para encontrar precisamente la recta que cumple con la condición de que la fórmula de mínimos cuadrados [](https://www.codecogs.com/eqnedit.php?latex=%20%5Csum_i%20%5By%20-%20(a%2Bbx)%5D%5E2%20#0) da el resultado mínimo posible. Se dice que esta recta minimiza las sumas de los cuadrados de las distancias de los puntos a la misma recta. Precisamente a esta recta la denominaremos recta de ajuste por mínimos cuadrados. Está probado que este método es uno de los mejores para representar un conjunto de puntos. El método de mínimos cuadrados es el método por defecto que utiliza el modelo de regresión lineal.

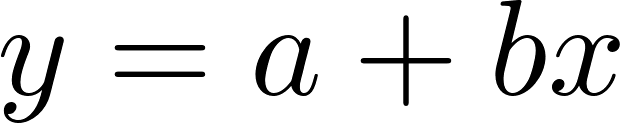
# El modelo de regresión lineal

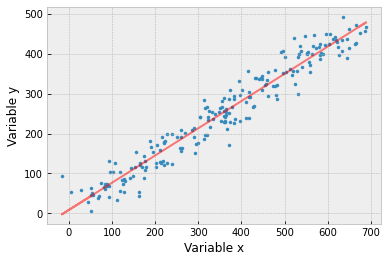
## El concepto del modelo

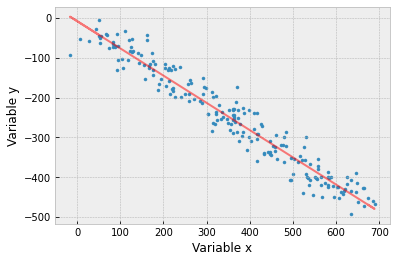
Por todo lo dicho, para poder aplicar un modelo de regresión lineal, debemos aplicar un ajuste por el método de mínimos cuadrados, y debemos hacer verificaciones para chequear que el modelo sea válido y bueno. Si esto se cumple, además de tener una recta de ajuste, la denominaremos recta de regresión, porque cumple con las condiciones que tiene que tener el modelo. Entonces es a partir de aquí cuando comenzamos a hablar de un modelo de regresión lineal, y no antes. Una vez validado el modelo, podemos usarlo para efectivamente representar a los datos y efectuar predicciones, como veremos más adelante.

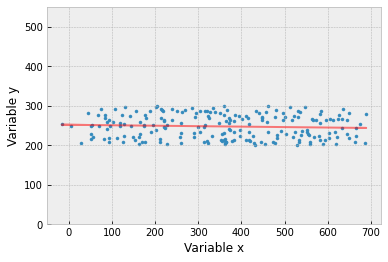
# Condiciones a cumplir

## El test de beta

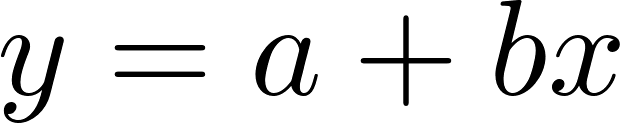
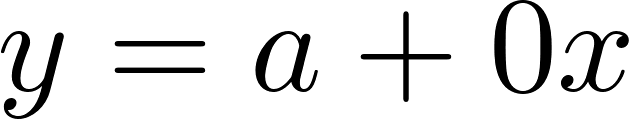
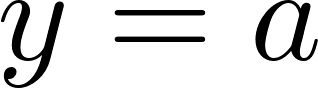
Como adelantábamos anteriormente, existen dos condiciones a cumplir para verificar la validez del modelo de regresión lineal. Por un lado, hablemos del “test de beta” o verificación de la pendiente, asociado al valor b de la función [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%20a%20%2B%20bx%20#0). La pregunta a responder es si efectivamente la recta es válida como representante del conjunto de datos. Veamos los siguientes casos:







Para el primer caso, donde hay una clara relación positiva, se ve que ante mayores valores de x, llegamos a mayores valores de y. En el segundo caso, es clara la situación pero aquí, dada la relación negativa, la variable y disminuye conforme x aumenta y viceversa. El tercer caso es el que queremos destacar para este tema. Aquí la recta parece tener una pendiente prácticamente nula, y la recta parece representar bastante apropiadamente los datos.

Pero tenemos un problema. Analicemos este caso desde un punto de vista puramente matemático. Supongamos, llevando este ejemplo al extremo, que la pendiente de b es efectivamente igual a cero. Entonces la función [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%20a%20%2B%20bx%20#0) se transforma en [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%20a%20%2B%200x%20#0), y por lo tanto el término que acompaña a la x, multiplicado por cero, se anula. Con lo que la función anterior queda como [](https://www.codecogs.com/eqnedit.php?latex=%20y%20%3D%20a%20#0). Vemos aquí que la x desapareció de la fórmula, y esto se traduce en el “desacoplamiento” de x e y. Esto significa que la variable y deja de ser dependiente de x. Como no tenemos dependencia, ya no podemos aplicar el modelo de regresión.

Este es un resultado muy importante. Veamos un resumen de lo planteado

* ¿Tienen los puntos una relación entre sí? Sí
* ¿Tiene la relación una forma “lineal”? Sí.
* ¿Existe una recta que pueda ser calculada con el método de mínimos cuadrados, y que en ese sentido pase por la parte “más central” de los datos”? Sí.
* ¿Puede armarse un modelo de regresión lineal a partir de este ajuste? No
* La recta de mínimos cuadrados, ¿representa adecuadamente a los datos? No
* ¿Por qué? Porque al ser la pendiente de la recta igual a cero, no hay dependencia de la variable y hacia la variable x. Y como no hay dependencia, no existe un modelo de regresión lineal que represente adecuadamente a estos datos.

En este caso el modelo de regresión lineal no sirve. ¿Cómo nos damos cuenta cuando pasa esto? Hablemos ahora del test de beta. Este es un test estadístico para verificar que la pendiente no sea cero. Puede ser positiva o negativa, pero no debería ser cero. No entraremos aquí en los detalles del test, pero sí aprenderemos a ver sus salidas. En Python podemos obtener esta información con el paquete extra pingouin. Las siguientes salidas fueron obtenidas con este paquete, y corresponden al primero y tercer caso planteados anteriormente.

| **Caso 1** | **names** | **coef** | **se** | **T** | **pval** | **r2** | **adj\_r2** | **CI[2.5%]** | **CI[97.5%]** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Intercept | 7.94 | 4.65 | 1.71 | 0.09 | 0.94 | 0.94 | -1.22 | 17.10 |
| **1** | x1 | 0.68 | 0.01 | 58.94 | 0.00 | 0.94 | 0.94 | 0.66 | 0.71 |
|  |  |  |  |  |  |  |  |  |  |
| **Caso 3** | **names** | **coef** | **se** | **T** | **pval** | **r2** | **adj\_r2** | **CI[2.5%]** | **CI[97.5%]** |
| **0** | Intercept | 251.70 | 4.53 | 55.59 | 0.00 | 0.01 | 0.0 | 242.77 | 260.63 |
| **1** | x1 | -0.01 | 0.01 | -1.02 | 0.31 | 0.01 | 0.0 | -0.03 | 0.01 |

Veamos aquí dos valores:

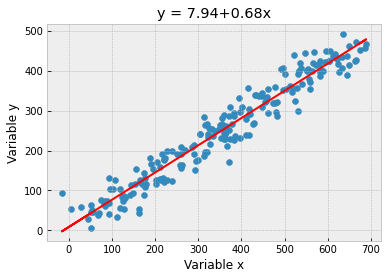
* En primer lugar, el “pval”, o p-value, que representa cuál es la probabilidad de que la recta en cuestión tenga valor cero. Esta probabilidad tiene que ser menor al 5%, por convención estadística. Por lo tanto, en números decimales debería ser un número menor a 0.05. Si esto ocurre (como en el caso 1), decimos que beta pasa el test, y por lo tanto la regresión es válida, o sea que la pendiente no es cero. En cambio, en el caso 3, vemos que la probabilidad es muy alta (0.31 = 31%), por lo que no podemos afirmar que la pendiente sea distinta que cero. Por lo tanto, aquí beta no pasa el test y la pendiente está tan cerca de cero que podemos considerar con rigurosidad estadística que el modelo de regresión lineal no sirve para ese conjunto de datos.
* En segundo lugar, con un poco más de información, tenemos el CI o intervalo de confianza. Este intervalo, con sus límites inferior (2.5%) y superior (97.5%), nos indica con una “confianza” o probabilidad del 95% dónde estará el valor de la pendiente. Esta afirmación tiene sustento estadístico y sirve como verificación suficiente para tomar decisiones sobre el modelo que estamos estudiando.   
  Para el caso 1, la pendiente estará entre los valores 242.77 y 260.63. Para el caso 2, la pendiente estará entre los valores -0.03 y 0.01. Esto quiere decir que, con un alto nivel de confianza, la pendiente estará en un intervalo que incluye al valor cero, esto es, bien podría ser cero con una probabilidad del 95%.   
  El intervalo de confianza proporciona el mismo resultado que el p-value, y además agrega la información acerca de dónde podría encontrarse el valor de la pendiente. Si el intervalo de confianza tiene valores positivos, quiere decir que la pendiente es positiva; si tiene valores negativos, quiere decir que la pendiente es negativa. Si tiene un valor positivo y uno negativo, quiere decir que cubre al cero, y por lo tanto aquí se mantiene la conclusión de que la regresión lineal no sirve para este conjunto de datos.

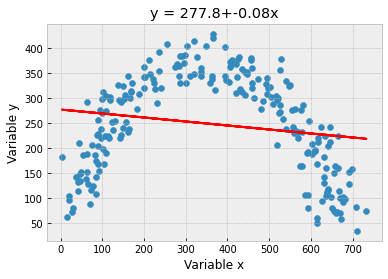
## El coeficiente de determinación

La segunda condición a cumplir es la del coeficiente de determinación. Este es un número que puede tomar valores entre 0 y 1, y puede interpretarse también de manera porcentual (de 0% a 100%). Se simboliza como R2 y representa el porcentaje de variabilidad de los datos explicada por el modelo de regresión lineal.

Veamos esta definición de forma detallada. La variabilidad es la cualidad que le da la forma al conjunto de datos. Cuando aplicamos por ejemplo el análisis de componentes principales, lo que estamos queriendo hacer es proyectar una “sombra” de los datos, y esta sombra se ve precisamente a partir de su variabilidad. La consecuencia de esta noción de variabilidad es que si podemos explicar o representar la variabilidad de los datos, podremos explicar su forma y así entenderlos mejor. Es precisamente esto lo que estamos haciendo con el modelo de regresión lineal. Por lo tanto, un modelo que funcione bien será el que mejor explique la variabilidad de los datos. Toda la variabilidad que no es explicada por los datos se debe al azar. Entonces el R2 será el grado porcentual de explicabilidad de los datos por parte del modelo, y su complemento (o sea lo que le falta en términos porcentuales para llegar a 100%) será lo que no puede explicar el modelo, o sea que queda en manos del azar.

Veamos los siguientes casos.





En el primer caso tenemos un conjunto de puntos que tiene una recta de ajuste que parece funcionar muy bien, por cuanto pasa bastante bien por el “centro” de los datos. Por su parte, en el segundo conjunto de datos no parece ser lo más conveniente utilizar una recta para representar los puntos. El valor de R2 aparece también en la salida de pengouin bajo la columna r2, con lo que el modelo del primer caso tiene un R2 del 94% (el modelo explica el 94% de la variabilidad de los datos), y el modelo del segundo caso tiene un R2 del 1%, con lo que por esta verificación puede descartarse como modelo poco útil. A continuación vemos los resultados con el valor de R2 resaltado

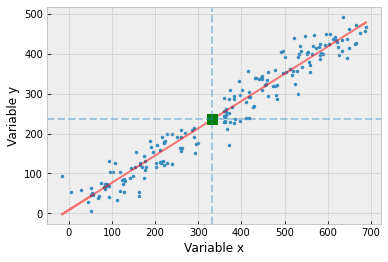
| **Caso 1** | **names** | **coef** | **se** | **T** | **pval** | **r2** | **adj\_r2** | **CI[2.5%]** | **CI[97.5%]** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Intercept | 7.94 | 4.65 | 1.71 | 0.09 | 0.94 | 0.94 | -1.22 | 17.10 |
| **1** | x1 | 0.68 | 0.01 | 58.94 | 0.00 | 0.94 | 0.94 | 0.66 | 0.71 |
|  |  |  |  |  |  |  |  |  |  |
| **Caso 2** | **names** | **coef** | **se** | **T** | **pval** | **r2** | **adj\_r2** | **CI[2.5%]** | **CI[97.5%]** |
| **0** | Intercept | 251.70 | 4.53 | 55.59 | 0.00 | 0.01 | 0.0 | 242.77 | 260.63 |
| **1** | x1 | -0.01 | 0.01 | -1.02 | 0.31 | 0.01 | 0.0 | -0.03 | 0.01 |

Curiosamente, como puede verse en la tabla, este último caso pasa satisfactoriamente el test de beta, con lo cual es un modelo válido, pero ciertamente poco útil. En este caso quizá convendría aplicar otro tipo de modelo que se ajuste a la forma curva de los datos. Tengamos entonces siempre en cuenta que el valor de R2 solamente tiene sentido una vez que se verificó y se pasó satisfactoriamente el test de beta mencionado anteriormente, y que además un modelo puede ser perfectamente válido pero no tener mucha utilidad por una baja capacidad de explicación de la variabilidad de los datos.

De esta forma, buscamos un modelo con un valor de R2 que sea aceptable, de tal forma que el azar, dado por su su complemento, sea lo más pequeño posible. Los valores aceptables de R2 tienen que ver con el campo de aplicación. En términos generales, para datos sociales y demográficos, podemos esperar valores de R2 superiores a 50% como aceptables, y para procesos físicos o químicos es deseable llegar a valores mayores a 90% como aceptables.

# Usar el modelo

Como mencionamos anteriormente, una vez validado el modelo en su test de beta y evaluada su calidad con el valor de R2, podemos darle uso. Esto consiste en aplicar el modelo para predecir valores desconocidos. En este caso, realizaremos predicciones para valores de y a partir de valores de x que no habían sido utilizados antes, utilizando la recta de regresión para hacer la predicción. El nuevo punto a predecir se colocará sobre la recta, y se podrá afirmar que el valor predicho tendrá una “fidelidad” igual al valor de R2. Por ejemplo, si hay algún valor de x faltante en el conjunto de datos, podemos suponer que su valor de y correspondiente corresponderá al señalado por la recta para dicho valor de x, como se muestra a continuación con el punto cuadrado de color verde.



En código de Python, veremos posteriormente que la aplicación es muy sencilla. Como todos los métodos de Data Science, el modelo se utiliza simplemente a través de la función predict().

FIN