**Los Modelo:** es una construcción conceptual simplificada para explicar una realidad más compleja. Ej: un Mapa que modela en 2D o 3D. Busca el equilibrio en aproximarse a la realidad y fácil de ser comprendido.

Cuando un modelo se vuelve muy complejo en el cual hay una cantidad tan grande de excepciones a tomar en cuenta que es complejo de explicar lo mejor es recurrir a la PROBABILIDAD: por ejemplo:

Es más fácil decir que el 97.7% de las ave pueden volar, y el restante es por fisiología, ambiente externo, ambiente interno, etc…

La probabilidad, nos ayuda a resumir nuestra incertidumbre sobre un tema, por falta de datos o pereza. Usando esto, da como resultado **Modelos Probabilísticos.**

La realidad vs lo que se trata de predecir con los datos, a veces los modelos no son exactos, y tiene un cierto error, donde si vemos el modelo del espacio de Aristóteles, Copérnico y Kepler cada uno reduce más el error a mínimo, el modelo del Sistema Solar.

**Los Datos,** el contacto que hacemos con la realidad, son multidimensionales, ya sabemos eso by BD relacionales o DE. Cada atributo es una dimensión.

**Parámetros,** se plantean restricciones, pero se presentan también ciertos valores que se pueden modificar, para ser flexibles y buscar hacer el modelo lo mas optimo posible.

**Machine Learning,** encontrar aquellos algoritmos que a través de los datos aprender cuales son los valores óptimos.

**El Error,** lo que no se mide no se mejora, una función de error nos diga como nuestro modelo se ajusta o no a los datos. Es a partir del cual se busca entrenar al modelo, o ajustarlo, es la **optimización** del mismo.

**En algoritmos de aprendizaje supervisado,** esta función se computa en los datos de salida suministrados.

**En algoritmos de aprendizaje no supervisado,** la función se computa en los datos de entrada.

La **Regresión Lineal**, es una técnica estadística, para predecir o estimar, una variable cuantitativa en función de otra variable cuantitativa.

La primera seria **Y**, la segunda seria **X**, la **Y** es la dependiente, ya que depende del valor que le demos a **X**, y la **X** es nuestra variable independiente, esta explica a **Y.**

**Y,** es aquella que necesitamos predecir o estimar

**X,** es nuestra variable explicativa

Consiste en modelar una ecuación de la recta, recordando cálculo diferencial, esta puede ser descendente o ascendente.

**Ejemplo:**

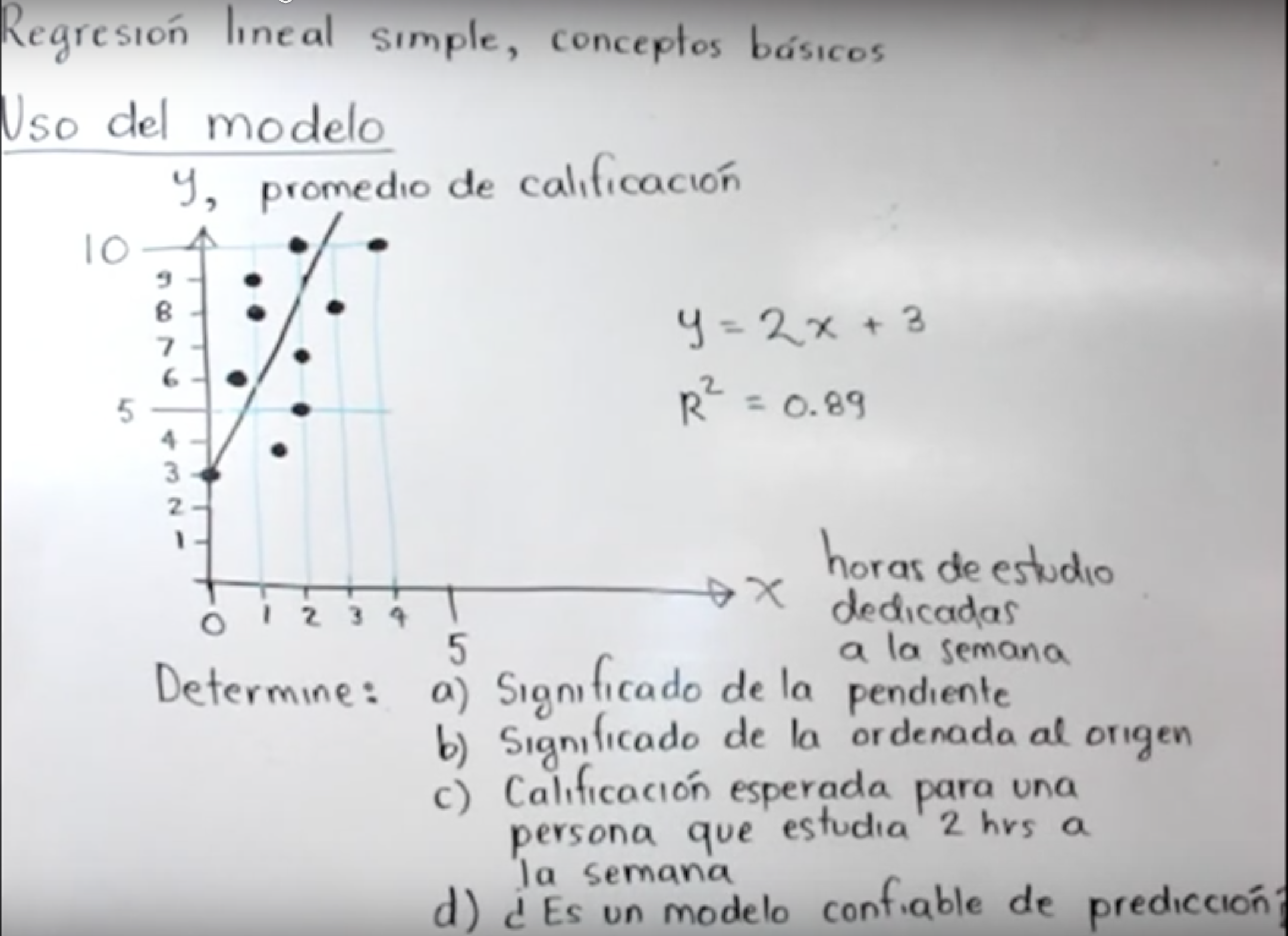
**Predecir el sueldo mensual en función a la edad de un trabajador**

**Y <- Sueldo** DEPENDIENTE

**X <- Edad** INDEPENDIENTE

Sii la edad sube (x), el sueldo sube (y), y viceversa…

**Conclusión:** Si ambas variables aumentan o disminuyen entonces tienen una relación directa.



1. **Significado de la pendiente**

Podemos decir que, si **X** son las horas de estudio y **Y,**  son el promedio de calificación obtenido, entonces por cada hora adicional o incremento unitario, la calificación aumenta dos puntos y = 2x + 3

1. **Significado de la ordenada al origen**

Que una persona que no estudie (0) horas, se espera que saque en promedio 3 puntos. ya que si 2 (0) + 3 = **3**

1. **Calificación esperada para una persona que estudia 2 horas a la semana**

Simplemente sustituimos, y ponemos Y = 2 (2) + 3 = **7**

1. **Es un modelo confiable de predicción?**

Se dice que el modelo es confiable ya que el coeficiente de determinación R^2, se acerca más a 1, ya que es 0.89

**Ejemplo:**

Deseamos predecir el tiempo que demora una constructora en construir un condominio en función al número de trabajadores que contrató de realizar la obra

**Y,** tiempo de construcción, a predecir

**X**, número de trabajadores, explicativa

**Conclusión:** Mientras más trabajadores contratan, menos tiempo se demorara el tiempo de construcción. Y mientras menos trabajadores, más tiempo se demorara el tiempo.

X y Y tienen una relación inversa, qué es lo contrario a una relación directa…

**Ejemplo:**

Predecir la nota de un alumno, en función a la cantidad de horas a la semana que estudio. Tenemos de muestra seis (6) alumnos:

|  |  |
| --- | --- |
| **Hrs de estudio (X)** | **Nota (Y)** |
| 3 | 8 |
| 6 | 10 |
| 8 | 15 |
| 2 | 8 |
| 1 | 5 |
| 6 | 12 |

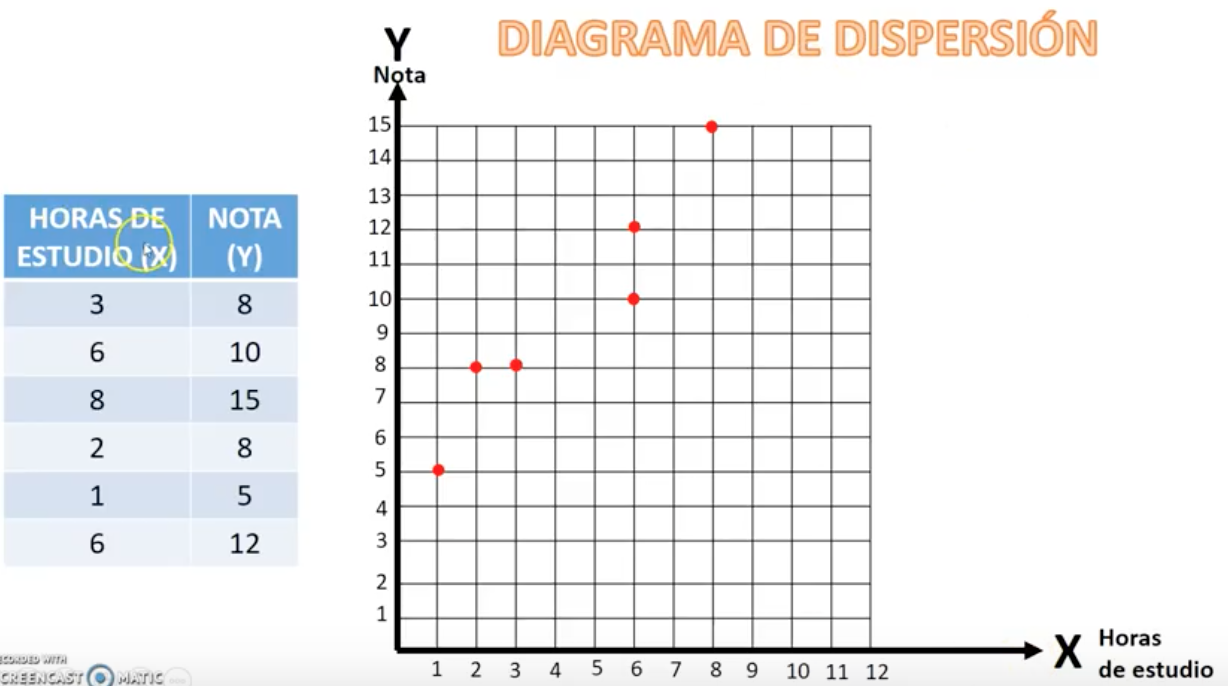
**IMPORTANTE:** Lo ideal a realizar, es un **diagrama de dispersión,** él mismo nos ayudará a ver la relación entre las variables **X** y **Y.**

**El Diagrama de dispersión,** no es más que graficar ambas variables en un plano cartesiano, mediante puntos de coordenadas.

X <- horas de estudio, explicativa

Y <- nota, a predecir

Se establece una **relación directa**, ya que a medida que la X (horas de estudio) aumenta, entonces la Nota (Y), va aumentando de igual manera...

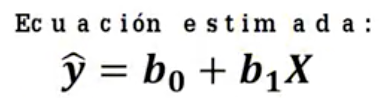


La técnica de regresión lineal, consiste en modelar una línea recta que une los puntos; no existe una recta como tal que pase por cada punto, deberían estar alineados los puntos…

Podemos decir que se puede intentar graficar, una línea donde se abarque la mayor cantidad de puntos. **Existen infinitas rectas posibles que pasan cerca de los 6 puntos.**

Pero como todo en la vida, existe una manera mas optima de hacerla, esta es la **Ecuación Estimada o ÿ (**y sombrero**)...**

La Ecuacion para sacar la línea recta, mediante la ecuación estimada:



**La mejor recta**, se define como aquella que logra disminuir las distancias entre los puntos y la recta como tal que se modela…

La Ecuación estimada Y, se obtiene calculado los coeficientes b0 y b1 con el **metodo de minimos cuadrados…**

Pero quienes son los coeficientes b0 y b1? quienes son estas constantes ???

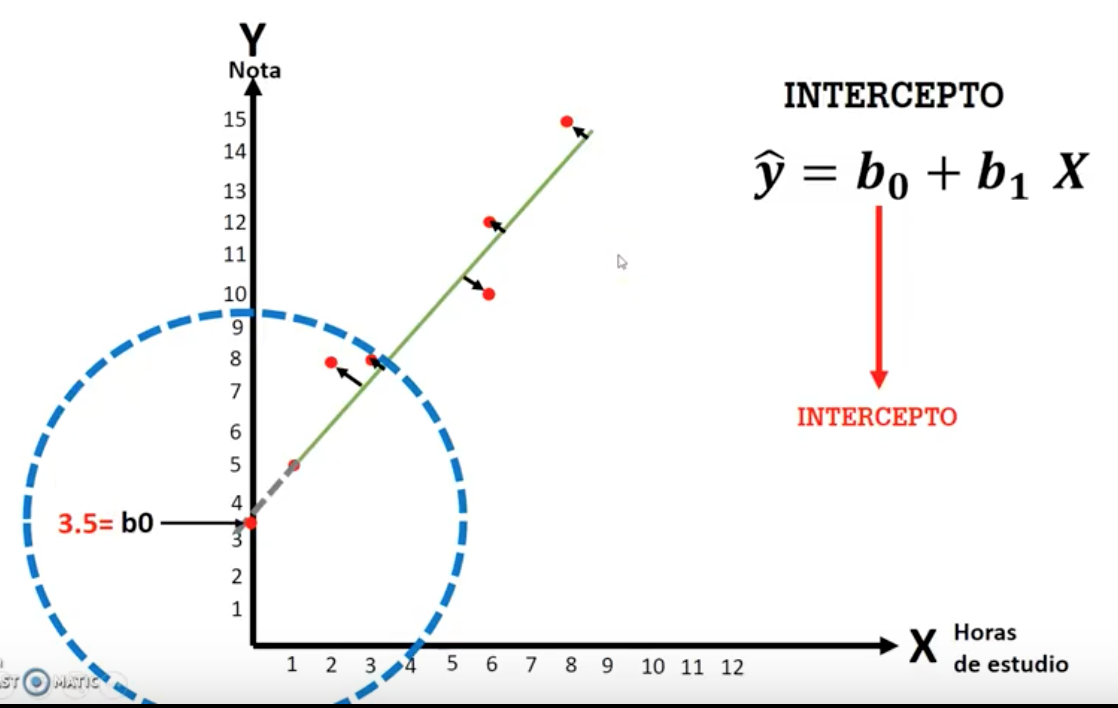
b0 se llama intercepto

b1 se llama pendiente

En la ecuación de la recta lo conocemos mejor como:

**Y = b + mX**

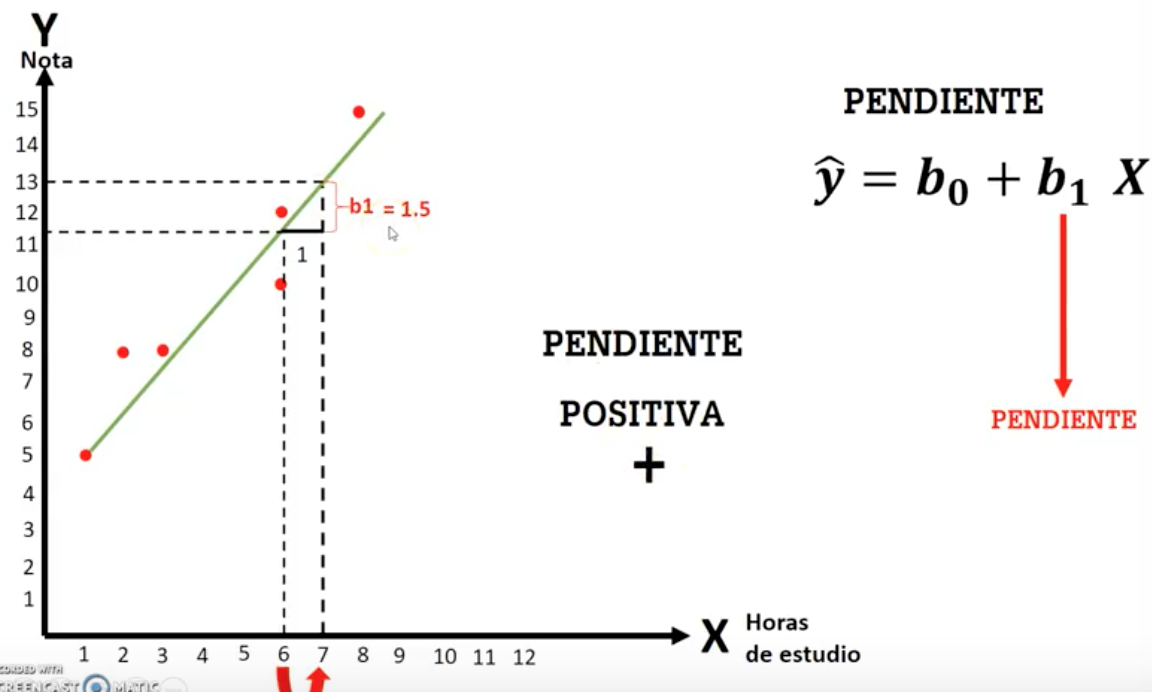
Donde b es el intercepto y m es la pendiente de la recta,

b0 o b, viene siendo el punto el cual donde la recta intercepta al eje **Y**

Si la recta no da como esa, se alarga un poco más,

Ahora bien la pendiente (m) o b1, indica la inclinación de nuestra recta con respecto al eje horizontal (X). A mayor pendiente, mayor es su valor, y más positiva, si la pendiente es hacia abajo, será negativa, horizontal es 0 y vertical totalmente se conoce como asíntota y es indeterminada.

En nuestra gráfica, la pendiente es positiva, escogemos en una distancia de 1 unidad, por ejemplo de **6 a 7,**  cualquier números elegidos funciona, el triángulo formado, la altura del mismo, es decir el cateto opuesto, es la pendiente, se mide hasta donde llega en el eje Y, y se calcula la distancia entre los puntos, y esa es nuestra pendiente m (b1):



La ecuación estimada es:

**Y = 3.5 + 1.5X**

Con este método, buscamos es estimar la variable Y utilizando esta ecuación a partir de la **X.**

Si estudia 0 hora = da 3.5 de nota

Si estudia 1 hora = día 5 de nota

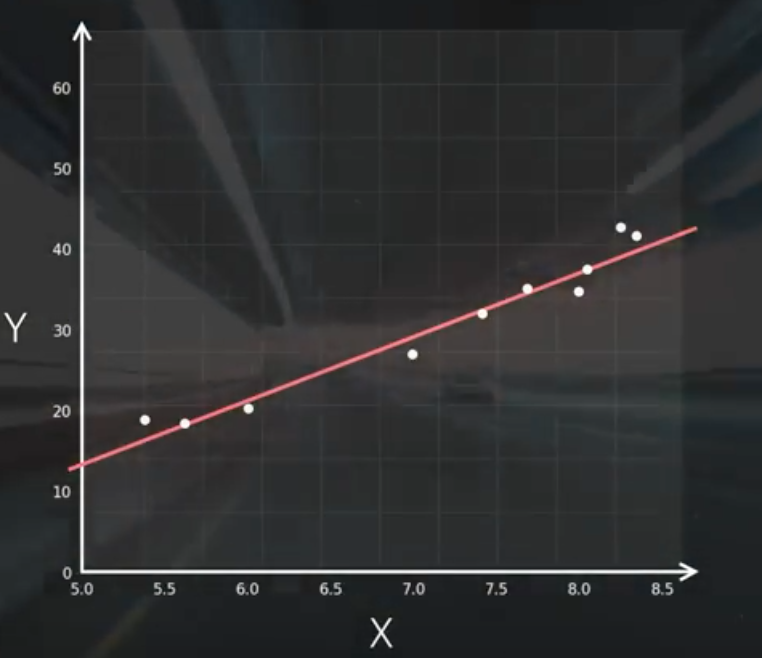
Si estudia 2 horas = da 6.5 de nota

Si estudia 4 horas = 9.5

Se nota un patrón y es que la nota aumenta en 1.5, que es nuestra pendiente

**Conclusión: por cada hora adicional de estudio, la nota del alumno aumenta en 1.5 puntos**

**From dot Csv**



Vamos buscando vivienda, y verificamos en distintos barrios de la ciudad. Notamos que:

**Cual es la relación entre el precio medio de la vivienda y el número medio de habitaciones por hogar?**

Barrio #1, promedio de hab es 6 y su precio medio son 20k USD,

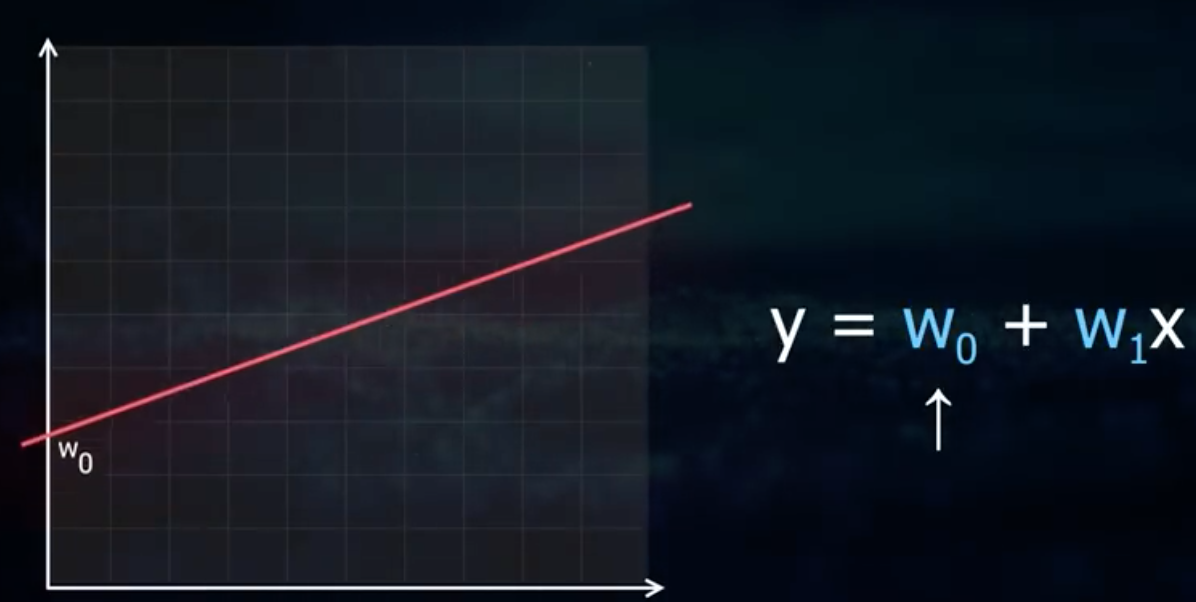
Barrio # 2, promedio de hab es 8 y su precio medio son 35k USD

Barrio # 3, promedio de hab es 7 y su precio medio es 28k USD

Nos damos cuenta que: a mayor numero de habitaciones, mayor cantidad de dinero mensualidad. Pues observando la grafica y los puntos, trazamos una linea recta entre ellos que represente esta **tendencia**, no debe de ser perfecta, y existe una **relacion directa,** Ya que si uno sube el otro tambien, si una variable baja, la otra tambien…

Es un Modelo… complejo en la realidad, pero facil de entender, que podemos predecir con el mismo. **Es el modelo de regresion lineal simple, 1 variable a predecir, y 1 explicativa**

Machine learning, que haga el modelo a partir de los datos solos...



**El intercepto**, a que altura corta el eje Y

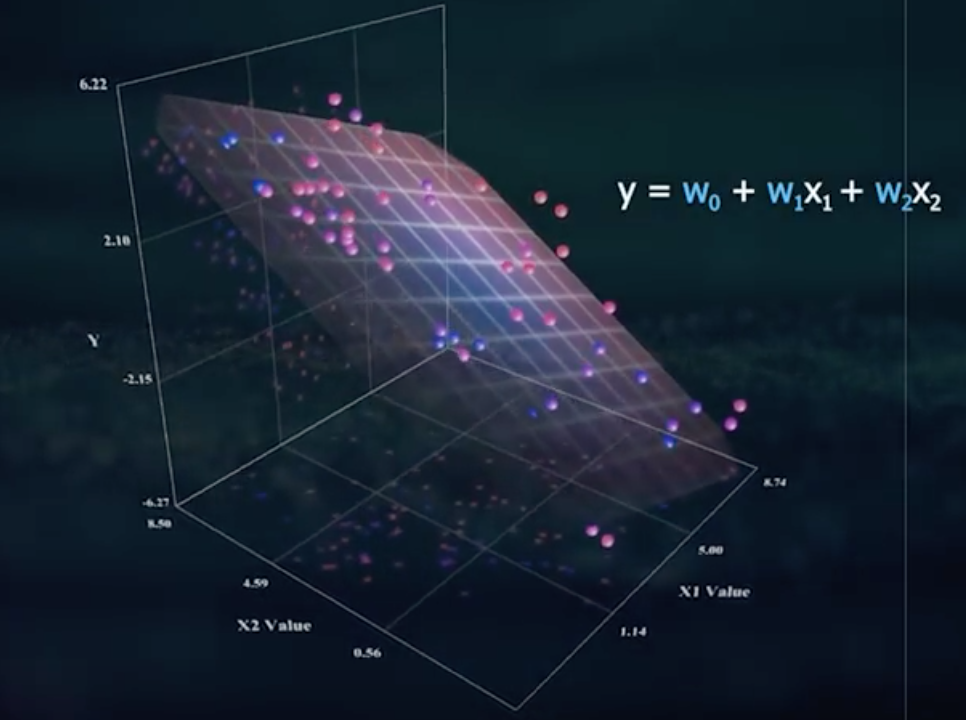
**La pendiente,** gráficamente la inclinacion de la recta y conceptualmente cual es la relacion en la variable de entrada X y la variable de salida Y

La regresion lineal simple, nos dice que solo una variable afecta a la otra, pero la realidad en la que vivimos no es asi…

En nuestra realidad, sabemos que multiples variables o factores influyen

1. La Cantidad de Habs
2. Cercanía de negocios
3. Grado de criminalidad, etc….

Ahora es un Modelo de Regresion Lineal Multiple…



Esto con dos (2) variables… Ya no sería bidimensional, sino tridimensional o multidimensional… **Hiperplanos**… **Cual seria el mejor plano que se adapte a nuestro modelo**...

Cada una de estas dimensiones representa una caracteristica de la realdiad que los datos representa, muchos atributos, muchas dimensiones. La forma mas facil ?

**Con vectores** (álgebra lineal)

Las **X**

Cada columna, una característica de los datos de entrada (hab, criminalidad...)

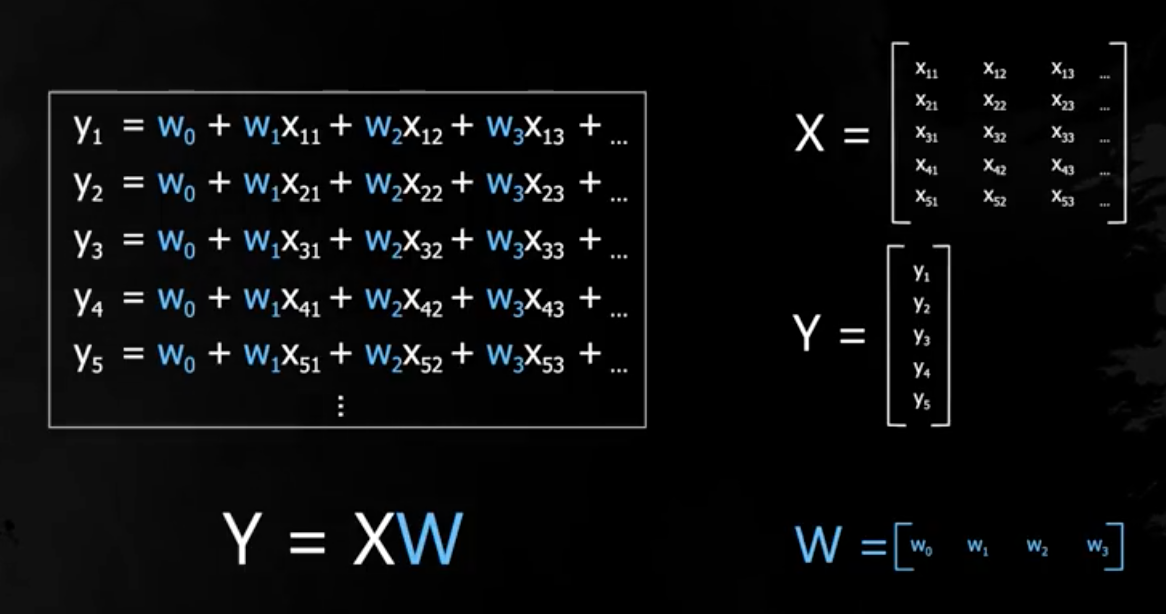
Cada fila, las mediciones de los conjuntos de datos (mediciones)

La **Y**

Seria un vector de elementos (la que queremos modelar)

La **W (**factores**)**

Nuestro vector de parámetros



**Y = XW**

A la hora de programar es mas facil y se entrene mucho mas eficientemente.

Volviendo con los modelos, lo mismo de la recta, cómo determinar la mejor

Para encontrar la recta que mejor represente los datos automaticamente, hay 2 metodos:

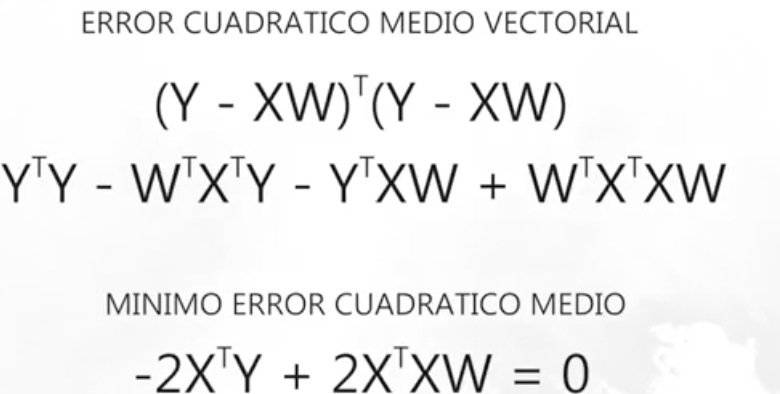
**Minimos Cuadrados Ordinarios,** la combinacion perfecta de parametros

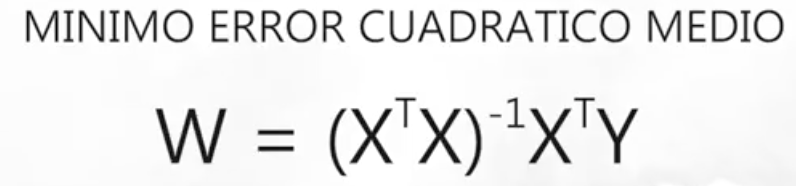
Recordemos que siempre hay un error, entre el valor que se predice y el valor real

Calcular la media de la funcion de coste: **mean((Yr - Ye)^2)** Error cuadratico medio, **la distancia entre el valor predicho y el valor real**, como no solo es un punto, sino que son muchos entonces se automatiza en una funcion, la **funcion de coste**

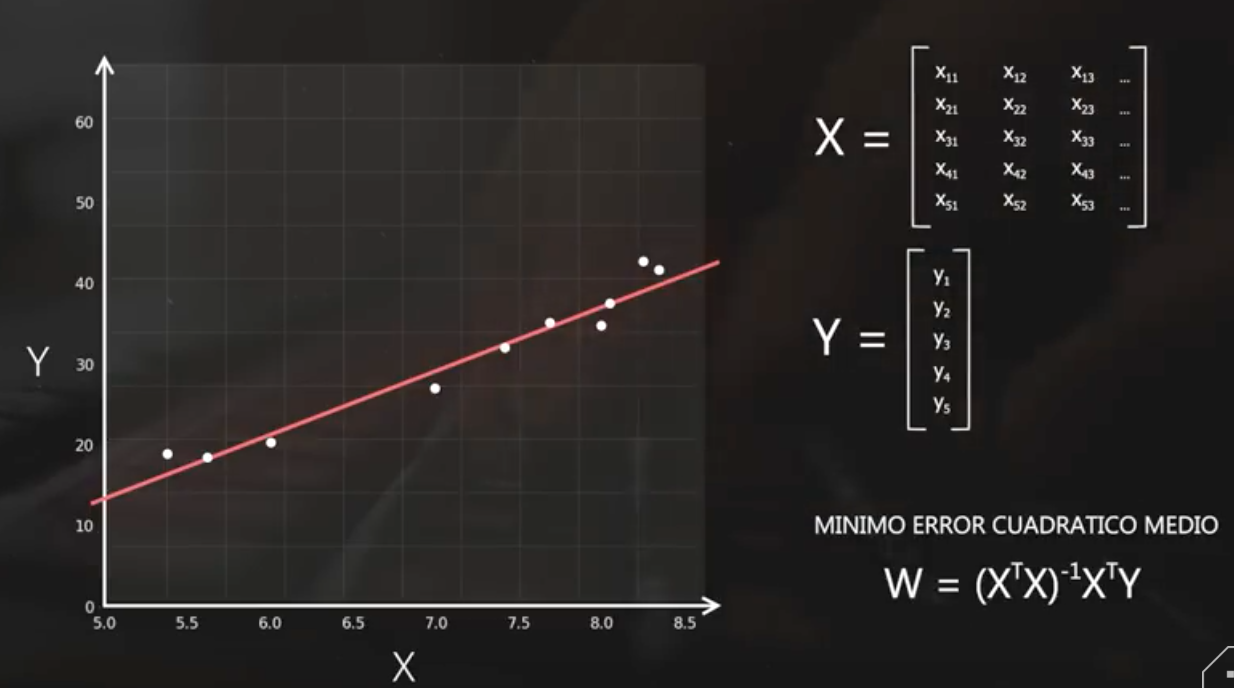
**Elevamos al cuadrado,** penalizando con mas intensidad los que estan mas lejos, y no afectando tanto a los que estan mas cerca. Recordando el valor minimo, la derivada de la funcion e igualandolo a cero

**Error cuadrático medio vectorial:** multiplicamos, y derivamos e igualamos a cero y despejamos para buscar el mínimo...





Minimizamos la funcion de coste original, de manera mas automatizada



No siempre se puede encontrar el minimo ya que puede volverse complejo

T es la traspuesta y -1 la inversa, Algebra lineal

El otro metodo:

**El Descenso del Gradiente…**

No se trata de una fórmula que nos da el valor directo que queremos, sino es iterativo que poco a poco nos acerca al **mínimo error.**