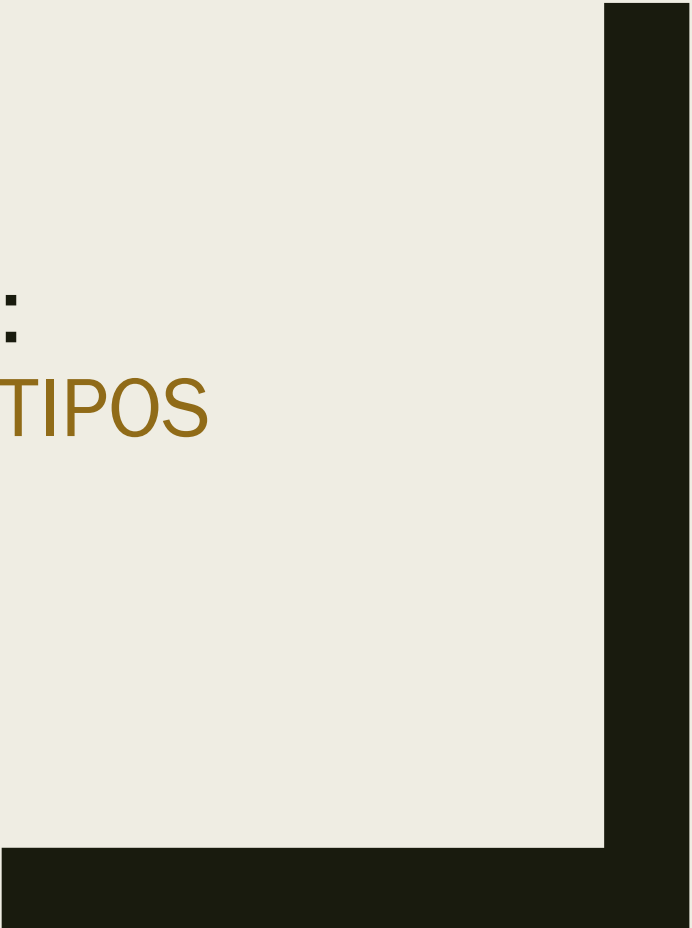




ANALÍTICA AVANZADA DE DATOS: MODELOS DE REGRESIÓN

A. Alejandra Sánchez Manilla
asanchezm.q@gmail.com



Regresión

- Es una técnica de análisis de datos sencilla, habitual y muy útil, conocida como "**ajustar una línea de tendencia**"
- Identifica la fuerza de la relación entre una o más características y una única etiqueta
- De naturaleza **paramétrica** porque hace ciertas suposiciones basadas en el conjunto de datos
- Su sencillez, comportamiento predecible, capacidad de previsión y alto nivel de interpretabilidad hacen que esta técnica se utilice en los ámbitos de las finanzas, la empresa, las ciencias sociales, la epidemiología y la medicina

Regresión

- También se le denomina como “ajustar una línea”
- En su forma más simple, la regresión ajusta una línea recta entre una variable (característica) y otra (etiqueta).
- En formas más complejas, la regresión puede encontrar relaciones no lineales entre una única etiqueta y múltiples características.

Puntos fuertes de la regresión

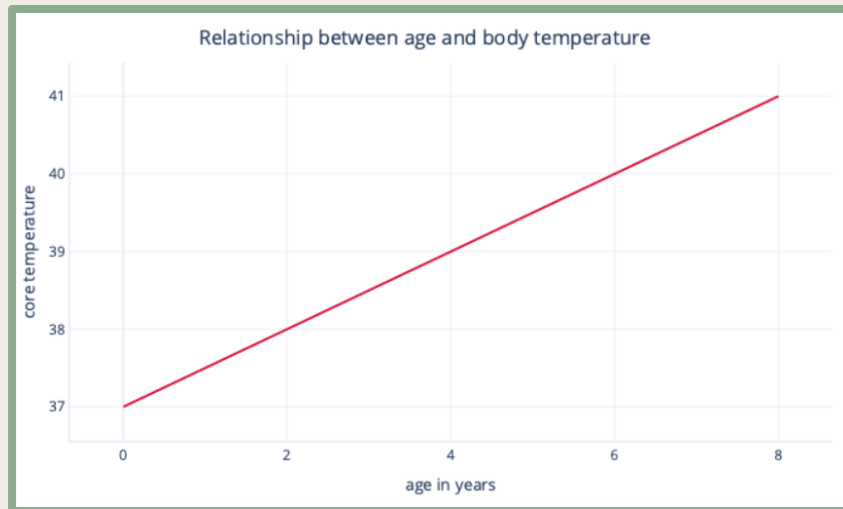
1. Predecibles y fáciles de interpretar:

- Porque describen ecuaciones matemáticas sencillas, que a menudo podemos representar gráficamente.
- De hecho, los modelos más complejos suelen denominarse soluciones “caja negra” (black box), porque es difícil entender cómo hacen predicciones o cómo se comportarán con determinadas entradas.

Puntos fuertes de la regresión

2. *Facilita la extrapolación:*

- Es decir, la predicción de valores fuera de rango de nuestro conjunto de datos.
- *Por ejemplo*, es sencillo estimar en la gráfica si un niño de nueve años tendrá una temperatura de 40.5°C .



NOTA:

¡Siempre hay que tener cuidado con la extrapolación!

Este modelo predeciría que una persona de 90 años tendría una temperatura casi tan alta como para hervir agua

Puntos fuertes de la regresión

3. *El ajuste óptimo suele estar garantizado*

La mayoría de los modelos de aprendizaje automático utilizan el descenso de gradiente para ajustar los modelos, lo que implica ajustar el algoritmo de descenso de gradiente y no garantiza que se encuentre una solución óptima.

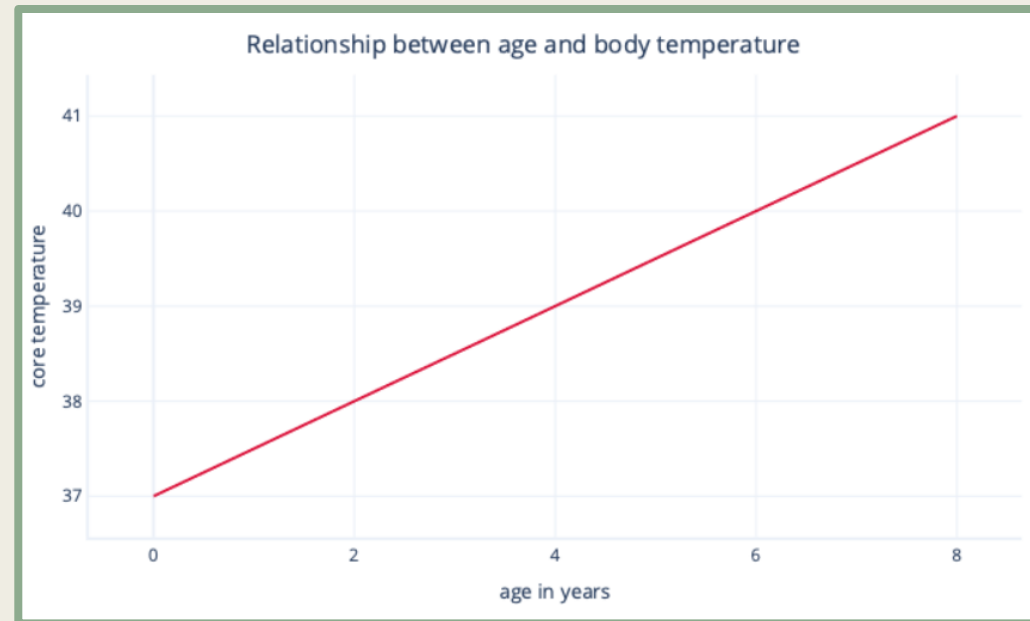
Con la *regresión lineal*, se utiliza la suma de cuadrados como **función de coste** no necesita un procedimiento iterativo de descenso de gradiente.

Se puede utilizar matemáticas inteligentes para calcular la ubicación óptima de la línea.

Es útil saber que (siempre que el tamaño de la muestra no sea demasiado grande) la regresión lineal no necesita que se preste especial atención al proceso de ajuste, y la solución óptima está garantizada

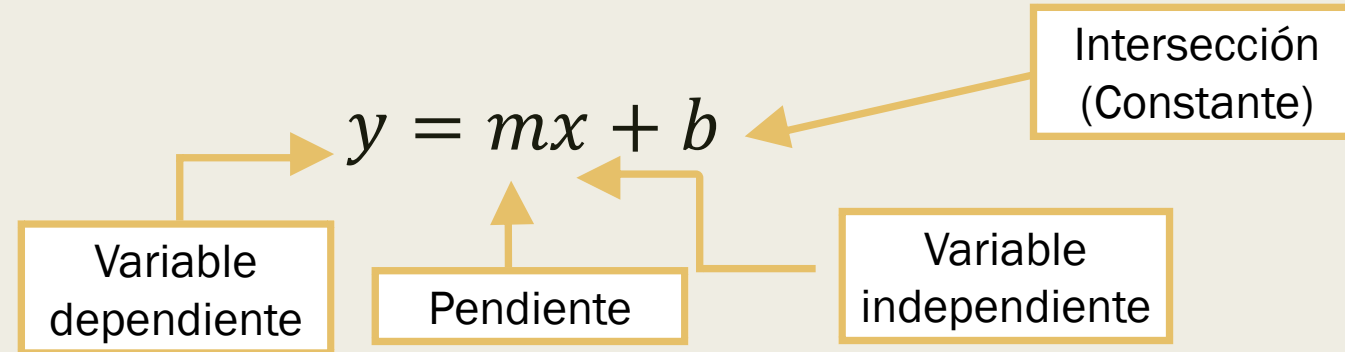
Regresión lineal simple

- Modela una relación lineal entre una única característica y una etiqueta, normalmente continua, lo que permite predecir la etiqueta a partir de características.
- Visualmente, puede tener este aspecto:



Regresión lineal simple

- Si lo pensamos matemáticamente, de manera sencilla sería:



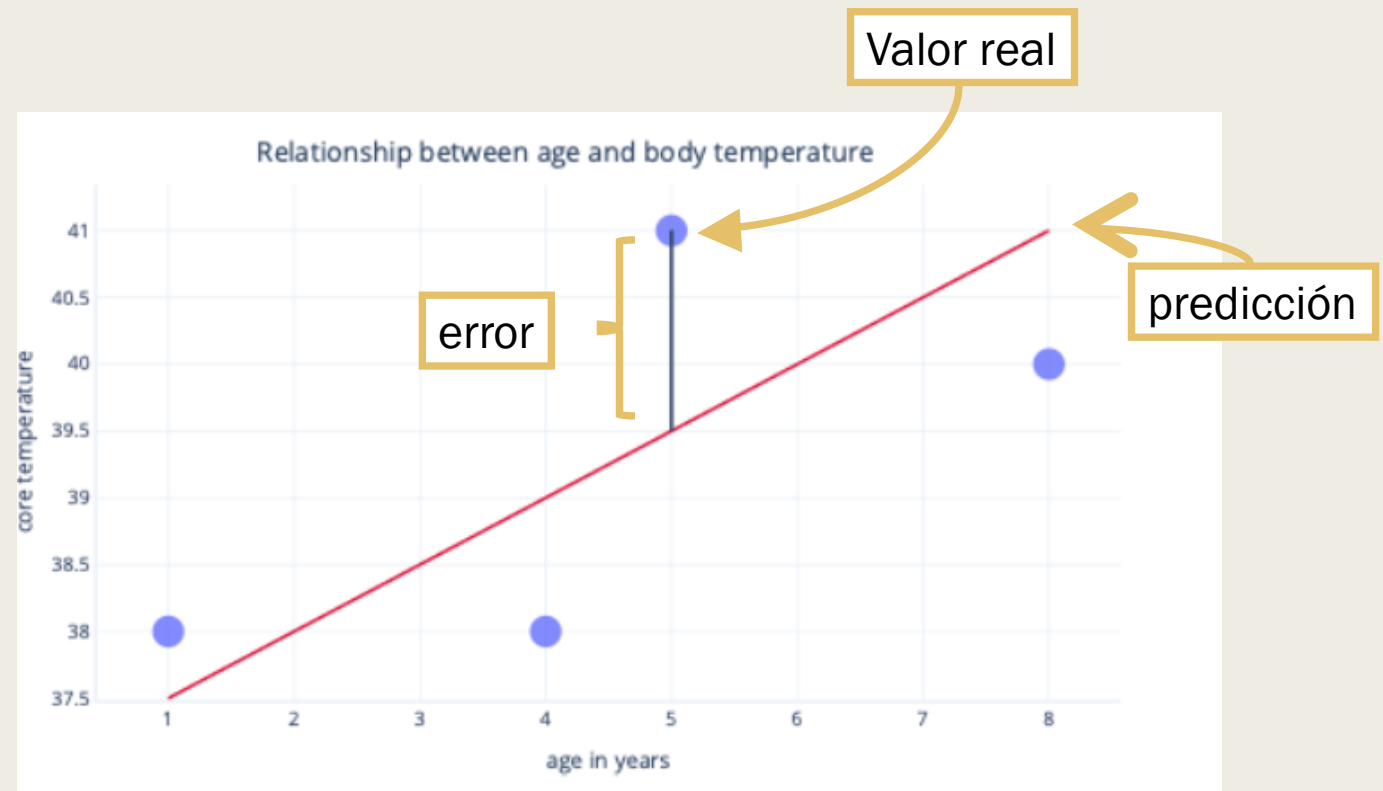
- ***Variables independientes o características (x)***, son variables que se manipulan para determinar el valor de la variable independiente; es decir, son las características que queremos usar para predecir algún valor dado de y .
- ***Variables dependientes u objeto (y)***, depende de los valores de la variable independiente; en otras palabras, es la característica que estamos tratando de predecir.

Regresión lineal simple

Ajuste de la regresión lineal

Objetivo:

Es encontrar la línea que produzca la menor cantidad de error, entendiendo por error la diferencia entre el valor real del punto de datos y el valor predicho



Regresión lineal simple

Error:

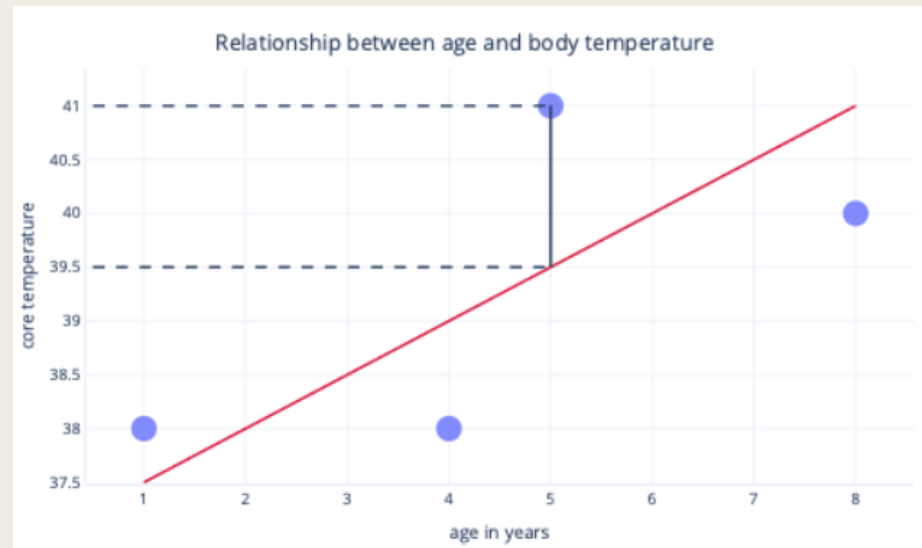
No importa cuán poderoso sea el algoritmo que elijamos, siempre habrá un error irreducible

No podemos eliminar por completo el error pero aún podemos intentar reducirlo al nivel más bajo

Regresión lineal simple

Ajuste de la regresión lineal

Retomando la gráfica anterior, si observamos estos dos puntos en un eje y , podemos ver que la predicción era 39.5, pero el valor real fue 41.



Por lo tanto, el modelo se equivocó en 1.5 para este dato.

Ajuste de la regresión lineal

Lo más habitual es ajustar un modelo minimizando la suma de cuadrados residuales. Esto significa que para la función de coste:

1. Calculando la diferencia entre los valores reales y los predichos para cada punto de datos
2. *Elevar al cuadrado estos valores*
3. *Suma (o promedio) de estos valores al cuadrado*

Este paso de elevar al cuadrado significa que no todos los puntos contribuyen por igual a la línea: *los valores atípicos*, que son puntos que no caen en el patrón esperado, tienen un error desproporcionadamente mayor, lo que puede influir en la posición de la línea.

Regresión lineal múltiple

- Modela la relación entre varias características y una única variable.
- Matemáticamente, es lo mismo que la regresión lineal simple y suele ajustarse utilizando la misma función de coste, pero con más características.
- En lugar de modelizar una única relación, esta técnica modela simultáneamente múltiples relaciones, que trata como independientes entre sí.
- Por ejemplo, si estamos prediciendo lo enfermo que se pone un perro en función de su edad y el porcentaje de grasa corporal, se encuentran dos relaciones
 - *Cómo la edad aumenta o disminuye la enfermedad*
 - *Cómo el porcentaje de grasa corporal aumenta o disminuye la enfermedad*

Regresión lineal múltiple

Si lo pensamos matemáticamente, sería:

The diagram shows the equation $y = a_1x_1 + a_2x_2 + \dots + a_nx_n + b$ with three callout boxes. A box labeled 'Variable dependiente' has an arrow pointing to 'y'. A box labeled 'Intersección' has an arrow pointing to 'b'. A box containing two lines of text has an arrow pointing to the coefficients 'a_1...a_n'.

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n + b$$

Variable dependiente

Intersección

$a_{1...n}$ Coeficientes
 $x_{1...n}$ Variables Independientes

Como podemos observar esta ecuación es muy parecida a la regresión lineal simple, solamente que incluimos las n variables independientes con su respectivo coeficiente

Por lo tanto, se manejan múltiples coeficiente y es computacionalmente más compleja por las variantes añadidas

Regresión lineal múltiple

Selección de variables independientes

No incluimos todas las variables independientes a la vez y posteriormente comenzamos a minimizar la función de error.



Seleccionar las mejores variables independientes que pueden contribuir a la variable dependiente

Construir una matriz de correlación para todas las variables independientes e incluimos la variable dependiente

Valor de correlación

Nos da una idea de qué variable es significativa y por qué factor

Regresión lineal múltiple

- A partir de esa matriz, elegimos las variables independientes en orden decreciente de valor de correlación y ejecutamos el modelo de regresión para estimar los coeficientes minimizando la función de error
- Nos detenemos cuando no hay mejora destacada en la función de estimación mediante la inclusión de la siguiente característica independiente

Este método aún puede complicarse cuando hay un gran número de características independientes que tienen una contribución significativa al decidir nuestra variable dependiente

Es importante tomar en cuenta con este método que agregar más variables independientes no significa que la regresión sea mejor u ofrece mejores predicciones

Regresión lineal múltiple

Matriz de correlación

- Una matriz de correlación es una tabla que indica los coeficientes de conexión entre los factores. Cada celda de la tabla muestra la conexión entre los dos factores.
- Es una tabla de doble entrada para A B y C, que muestra una lista multivariable horizontalmente y la misma lista verticalmente y con el correspondiente coeficiente de correlación llamado r o la relación entre cada pareja en cada celda, expresada con un número que va desde 0 a 1. El modelo mide y muestra la interdependencia en relaciones asociadas o entre cada pareja de variables y todas al mismo tiempo.

Variables	A	B	C
A	1	0,3	0,75
B	0,3	1	0,95
C	0,75	0,95	1

La mejor relación es B C o C B (0.95) ya es alta.

La diagonal de 1 -unos- no tiene significado, únicamente forma una línea divisoria entre valores que se repiten a ambos lados como en un espejo.

Regresión lineal múltiple

La regresión lineal múltiple tiene supuestos

- El hecho que el modelo espere que las características sean independientes se denomina **suposición del modelo**. Cuando las suposiciones del modelo no son ciertas, el modelo puede hacer **predicciones engañosas**.

Ejemplo:

- La edad probablemente predice cómo enferman los perros, ya que los perros mayores enferman más
- Junto con el hecho de si a los perros se les ha enseñado a jugar al frisbee: probablemente todos los perros mayores saben jugar al frisbee (*saber_frisbee*).

La regresión lineal múltiple tiene supuestos

Si incluyéramos la edad y *saberfrisbee* en nuestro modelo como características, probablemente nos diría:

- Que *saberfrisbee* es un buen predictor de una enfermedad y subestimaría la importancia de la edad.

Suena absurdo porque:

- No es probable que saber jugar al frisbee provoque una enfermedad.
- Por el contrario, la raza (*raza_perro*) también podría ser un buen predictor de enfermedad, pero no hay razón para creer que la edad predice *raza_perro*, por lo que sería seguro incluir ambos en un modelo.

Regresión lineal múltiple

En algunos casos, agregar más variables independientes puede empeorar las cosas

Ajuste excesivo

Cuando se agrega más variables independientes se crean relaciones entre ellas

No solo las variables independientes están potencialmente relacionadas con las variables dependientes, sino que también están potencialmente relacionadas entre sí

Multicolinealidad

El escenario **óptimo** es que todas las variables independientes se correlacionen con la variable dependiente, pero no entre sí

Regresión lineal múltiple

¿Qué es el R^2 o coeficiente de determinación?

- Es una medida estadística que examina cómo las diferencias en una variable pueden ser explicadas por la diferencia en una segunda variable, al predecir el resultado de un evento determinado.
- Este coeficiente, que se conoce más comúnmente como R-cuadrado (o R^2), evalúa la fuerza de la relación lineal entre dos variables, y es muy utilizado por los investigadores cuando realizan análisis de tendencias
- Por ejemplo, este coeficiente puede contemplar la siguiente pregunta: si una mujer se queda embarazada un día determinado, ¿cuál es la probabilidad de que dé a luz en una fecha concreta en el futuro? En este escenario, esta métrica pretende calcular la correlación entre dos acontecimientos relacionados: la concepción y el nacimiento.

Regresión lineal múltiple

Bondad de ajuste R^2

- Sabemos que las funciones de coste pueden utilizarse para evaluar lo bien que un modelo se ajusta a los datos con los que ha entrenado
- Los modelos de regresión lineal tienen una medida especial relacionada llamada R^2 (“r-squared”)
- R^2 es un valor entre 0 y 1 que nos indica lo bien que un modelo de regresión lineal se ajusta a los datos
- Cuando se dice que las correlaciones son fuertes, a menudo se requiere decir que el valor R^2 es grande

Regresión lineal múltiple

Bondad de ajuste R^2

Los valores de R^2 son ampliamente aceptados, pero no son una medida perfecta que podamos utilizar de forma aislada. Sufren cuatro limitaciones:

Debido a cómo se calcula R^2 , cuantas más muestras tengamos, mayor será R^2 . Esto puede llevarnos a pensar que un modelo es mejor que otro, simplemente porque los valores de R^2 se calcularon utilizando diferentes cantidades de datos

Los valores de R^2 no nos dicen lo bien que funcionará un modelo con datos nuevos. Los estadísticos solucionan esto calculando una medida suplementaria, llamada valor p .

Regresión lineal múltiple

Bondad de ajuste R^2

Los valores de R^2 son ampliamente aceptados, pero no son una medida perfecta que podamos utilizar de forma aislada. Sufren cuatro limitaciones:

Debido a cómo se calcula R^2 , cuantas más muestras tengamos, mayor será R^2 . Esto puede llevarnos a pensar que un modelo es mejor que otro, simplemente porque los valores de R^2 se calcularon utilizando diferentes cantidades de datos

Los valores de R^2 no nos dicen lo bien que funcionará un modelo con datos nuevos. Los estadísticos solucionan esto calculando una medida suplementaria, llamada valor p .

Regresión polinomial

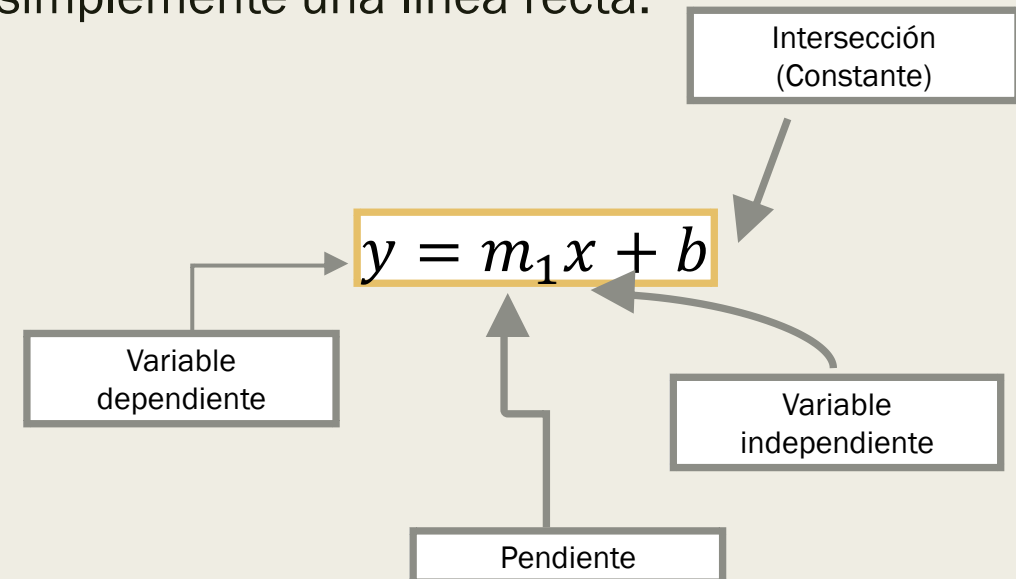
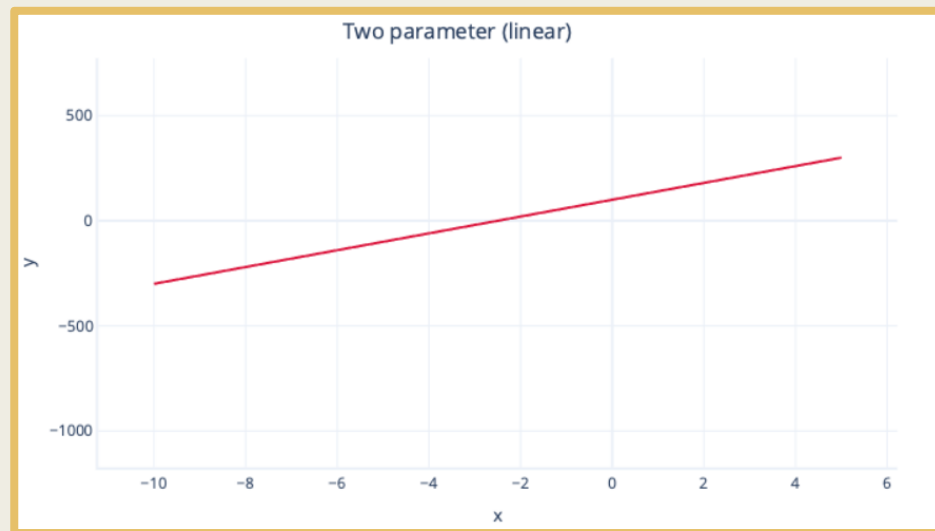
- Hasta ahora, sólo hemos visto modelos de regresión lineal, es decir, modelos que pueden modelarse como líneas rectas. Sin embargo, los modelos de regresión pueden funcionar prácticamente con cualquier otro tipo de relación.
- La **regresión polinomial** modela las relaciones como un tipo particular de curva. Los polinomios son una familia de curvas que van de formas simples a complejas. Cuantos más parámetros tenga la ecuación (modelo), más compleja puede ser la curva.
- Es decir, extiende el modelo lineal al agregar predictores adicionales, que se obtienen al elevar cada uno de los predictores originales a una potencia. Por ejemplo, una regresión cúbica utiliza tres variables independientes, como predictores.
- Este enfoque proporciona una forma sencilla de proporcionar un ajuste no lineal a los datos.

Regresión polinomial

- El método estándar para extender la Regresión Lineal a una relación no lineal entre las variables dependientes e independientes, ha sido reemplazar el modelo lineal con una función polinomial.

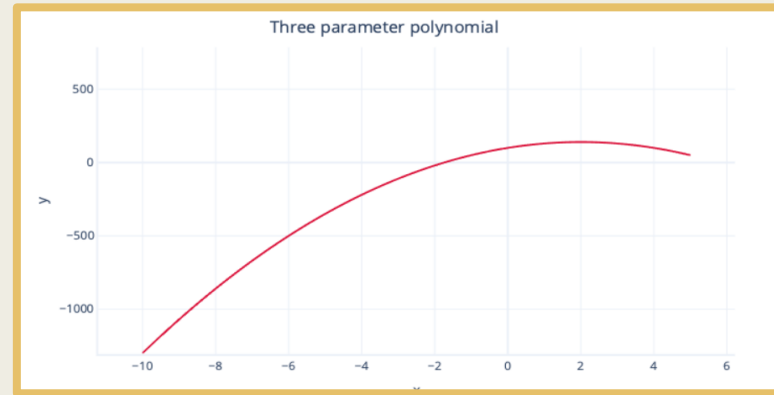
Este algoritmo es un ensayo y error en donde se debe probar con distintos grados de polinomios para obtener el que mejor se adecue a los datos.

- Por ejemplo, un polinomio de dos parámetros es simplemente una línea recta:



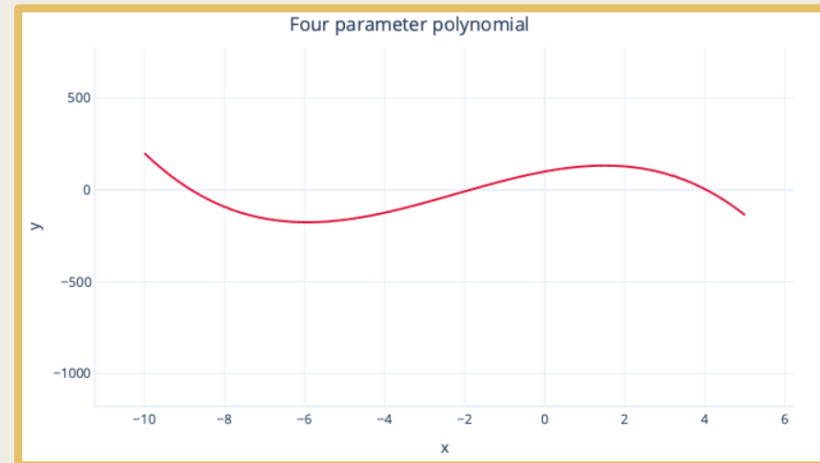
Regresión polinomial

- Mientras que un polinomio de tres parámetros tiene una sola curva:



$$y = m_1x + m_2x^2 + b$$

- Y un polinomio de cuatro parámetros puede tener dos curvas:



$$y = m_1x + m_2x^2 + m_3x^3 + b$$

Regresión polinomial

Regresión lineal simple

Con 1 variable

$$y = m_1x_1 + b$$

Regresión polinomial

$$y = m_1x_1 + m_2x_1^2 + b$$

Con 2 variable

$$y = m_1x_1 + m_2x_2 + b$$

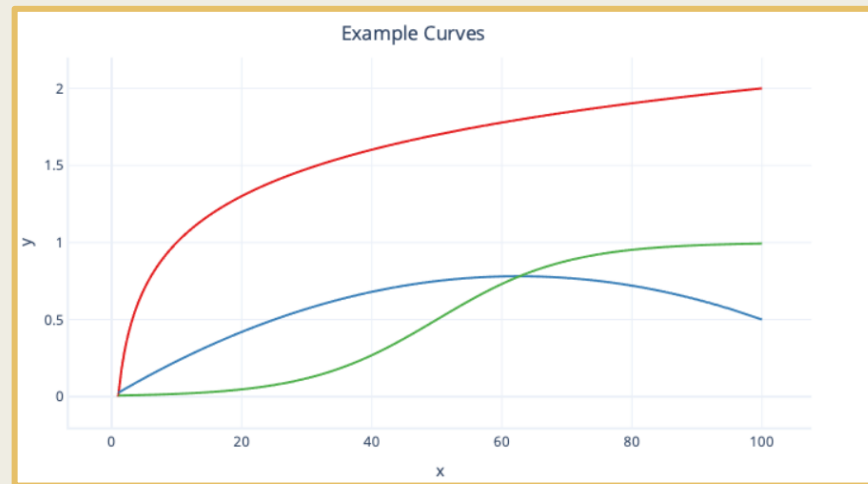
$$y = m_1x_1 + m_2x_1^2 + m_3x_2 + m_4x_2^2 + m_5x_1x_2 + b$$

Regresión polinomial

- Para la *regresión polinomial* se crean características adicionales que no se encuentran en la Regresión Lineal.
- Un término polinomial, bien sea cuadrático o cúbico, convierte un modelo de Regresión Lineal en una curva, pero como los datos de x son cuadráticos o cúbicos pero el coeficiente b no lo es, todavía se califican como un modelo lineal.
- Esto hace que sea una forma agradable y directa de modelar curvas sin tener que crear modelos complicados no lineales.

Curvas polinomiales frente a otras curvas

- Hay muchos tipos de curvas, como las curvas logarítmicas y las curvas logísticas (en forma de s), todas se pueden utilizar con la regresión.



- Una gran ventaja de la **regresión polinomial** es que puede utilizarse para analizar todo tipo de relaciones.

Curvas polinomiales frente a otras curvas

- Por ejemplo, la **regresión polinomial** puede utilizarse:
 - *Para relaciones que son negativas dentro de un determinado rango de valores de características, pero positivas dentro de otros.*
 - *Cuando la etiqueta (valor y) no tiene un límite superior teórico.*



Desventajas:

- A menudo extrapolan mal. En otras palabras, si intentamos predecir valores mayores o menores que nuestros datos de entrenamiento, los polinomios pueden predecir valores extremos poco realistas.
- Las curvas polinomiales son fáciles de **sobreajustar**. Esto significa que el ruido en los datos puede cambiar la forma de la curva mucho más que los modelos más sencillos, como la regresión lineal simple.

Tarea 2.

- Crear por equipo tres mapas mentales, uno por cada modelo de regresión visto en esta presentación:
 - *Regresión lineal simple*
 - *Regresión lineal múltiple*
 - *Regresión polinomial*

Tip: https://www.canva.com/es_mx/aprende/que-son-los-mapas-mentales/

Nota: No olviden subirlos a su carpeta DRIVE.