Regresión Lineal

Delta Analytics construye capacidad técnica alrededor del mundo.

El contenido de este curso está siendo desarrollado activamente por Delta Analytics, una organización sin fines de lucro 501(c)3 del Área de la Bahía que apunta a capacitar a las comunidades para aprovechar sus datos.

Por favor comuníquese con cualquier pregunta o comentario a inquiry@deltanalytics.org.

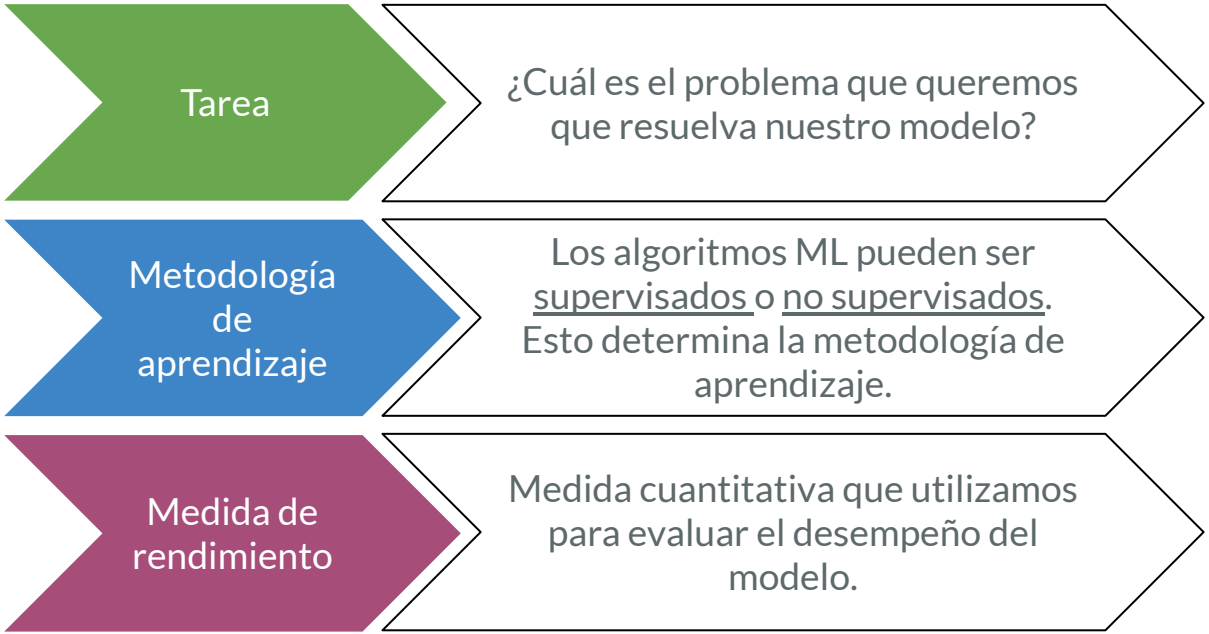
Descubre más sobre nuestra misión aquí.

Módulo 3:
Regresión Lineal

Hagamos una rápida revisión del módulo 2!

¿Cómo aprende un modelo desde los datos?

Todos los modelos tienen los siguientes componentes:



Fuente: Deep Learning Book - Chapter 5: Introduction to Machine Learning

Ejercicio 1

¿Cuáles son las características explicativas? ¿Cuál es la característica de resultado?

Tareas de clasificación:

Deuda pendiente	Ingreso anual actual ($)	Aprobación de tarjeta de crédito	Aprobación prevista para tarjeta de crédito
		Y	Y*
200	12,000	No	Yes
60,000	60,000	No	No
0	11,000	Yes	No
10,000	200,000	Yes	Yes

Completemos los espacios en blanco para nuestro ejemplo de aprobación de crédito:

Tarea	
Medida de Rendimiento	
Experiencia de aprendizaje	

Completemos los espacios en blanco para nuestro ejemplo de aprobación de crédito:

Tarea	¿Debe este cliente ser aprobado para una tarjeta de crédito?
Medida de Rendimiento	Pérdida Log
Experiencia de aprendizaje	Supervisado

Ejercicio 2

¿Cuáles son las características explicativas?
¿Cuál es la característica de resultado?

Tarea de regresión:

Tiempo usado: Una semana estudiando machine learning	Exactitud del modelo de clasificación construido por el alumno	Precisión prevista del modelo de clasificación
X	Y	Y*
10	90%	30%
2	30%	60%
12	95%	26%
0	50%	88%

Completemos los espacios en blanco para nuestro ejemplo de estudio:

Tarea	
Medida de Rendimiento	
Experiencia de aprendizaje	

Completemos los espacios en blanco para nuestro ejemplo de estudio:

Tarea	¿Qué tan preciso es el modelo de clasificación de este estudiante?
Medida de Rendimiento	MSE
Experiencia de aprendizaje	Supervisado

Módulo 3:
Regresión Lineal

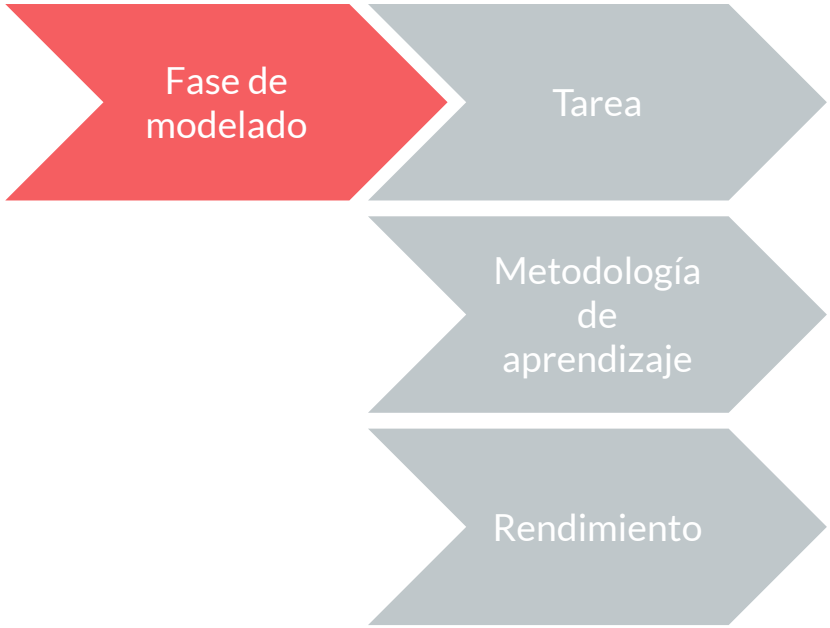
Module Checklist

- ❏ Regresión Lineal
  - ❏ Relación entre dos variables (x e y)
    - ❏ Formalizando f(x)
    - ❏ Correlación entre dos variables
    - ❏ Supuestos
  - ❏ Ingeniería de características y selección
  - ❏ Proceso de aprendizaje: función de pérdida y Error Cuadrático Medio
  - ❏ Regresión univariante, Regresión multivariada
  - ❏ Medidas de rendimiento (R2, R2 ajustado, MSE)
  - ❏ Overfitting, Underfitting (sobre-ajuste, sub-ajuste)

¿Qué es regresión lineal?

La regresión lineal es un modelo que explica la relación entre las características explicativas y una característica de resultado como una **línea en un espacio bidimensional.**

¿Por qué es importante la regresión lineal?



La regresión lineal ha estado en uso desde el siglo XIX y es **una de las herramientas de aprendizaje automático más importantes** disponibles para los investigadores.

Muchos otros modelos se basan en la lógica de los modelos lineales. Por ejemplo, la forma más simple de modelo de aprendizaje profundo, sin capas ocultas, es un modelo lineal.

Source: History of Linear Regression - Introduction to Linear Regression

Regresión lineal:

Pros

- Modelo muy popular con resultados intuitivos y fáciles de entender.
- Extensión natural del análisis de correlación.

Cons

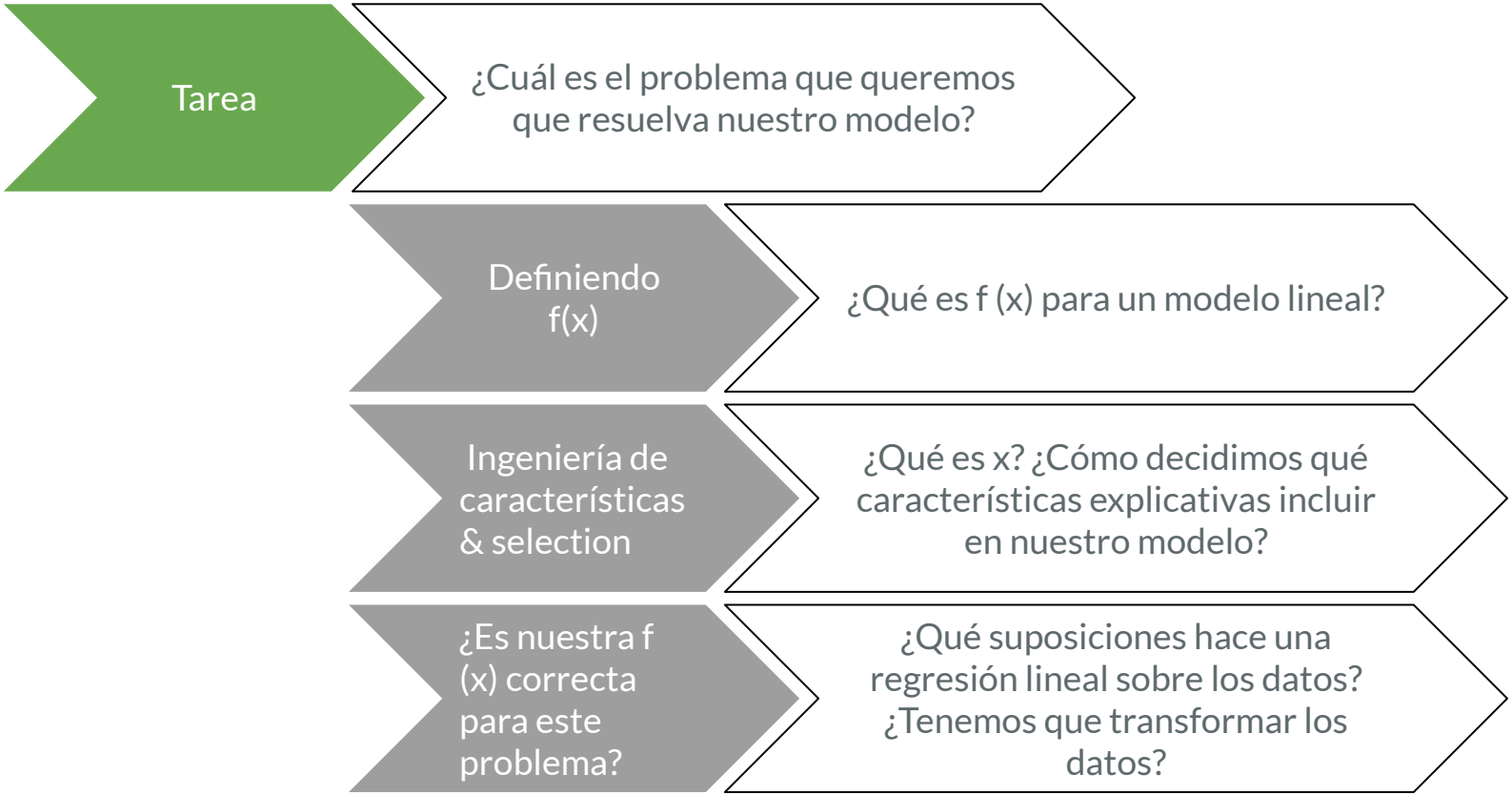
- Sensible a los valores atípicos
- El mundo no siempre es lineal; a menudo queremos modelar relaciones más complejas
- No nos permite modelar interacciones entre características explicativas (podremos hacerlo usando un árbol de decisión)

Supuestos*

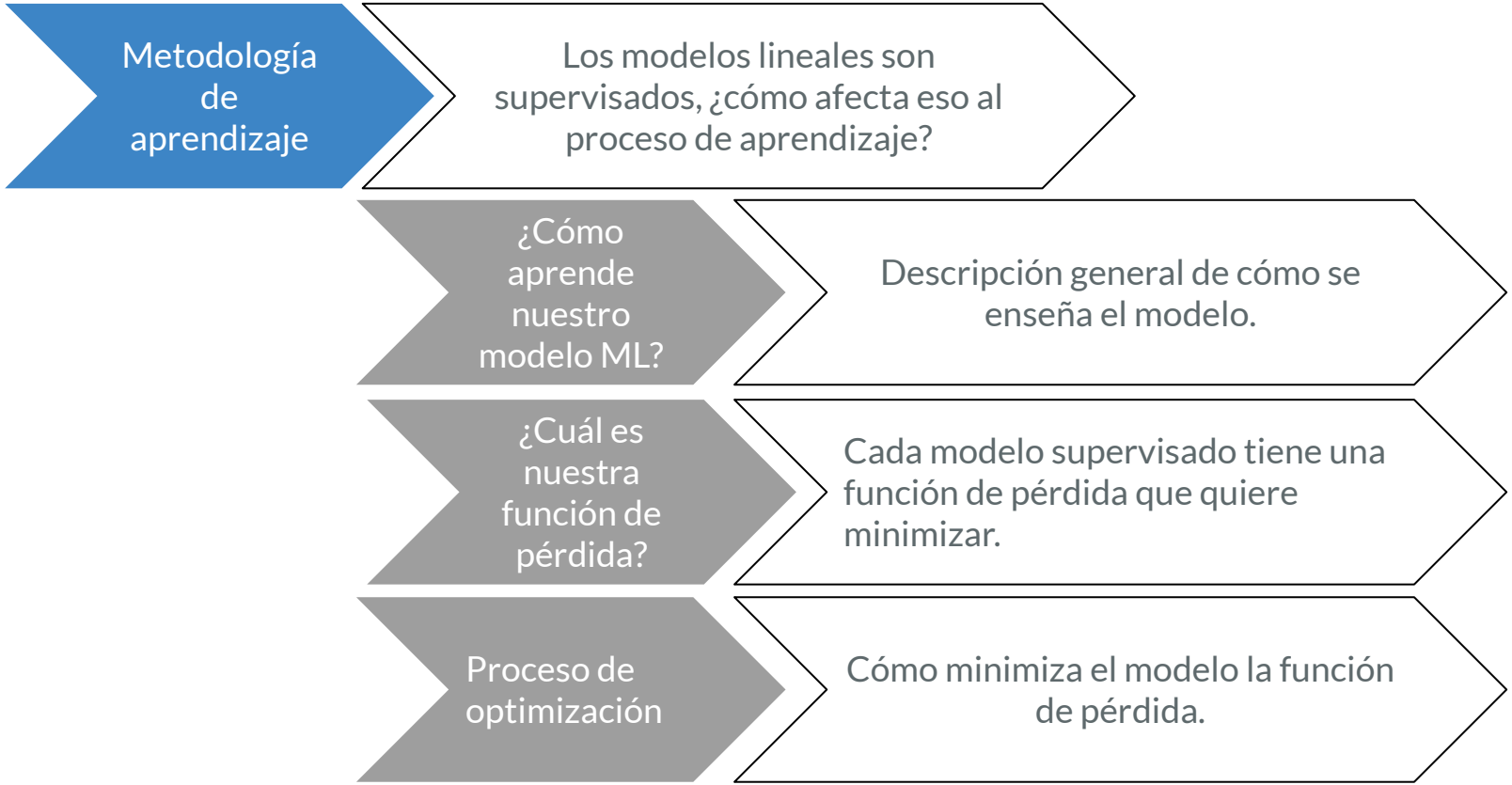
- Relación lineal entre x e y
- Distribución normal de variables.
- Sin multicolinealidad (variables independientes)
- Homocedasticidad
- *Regla general:* al menos 20 observaciones por variable independiente en el análisis

* ¡Explicaremos cada uno de estos supuestos en este módulo!

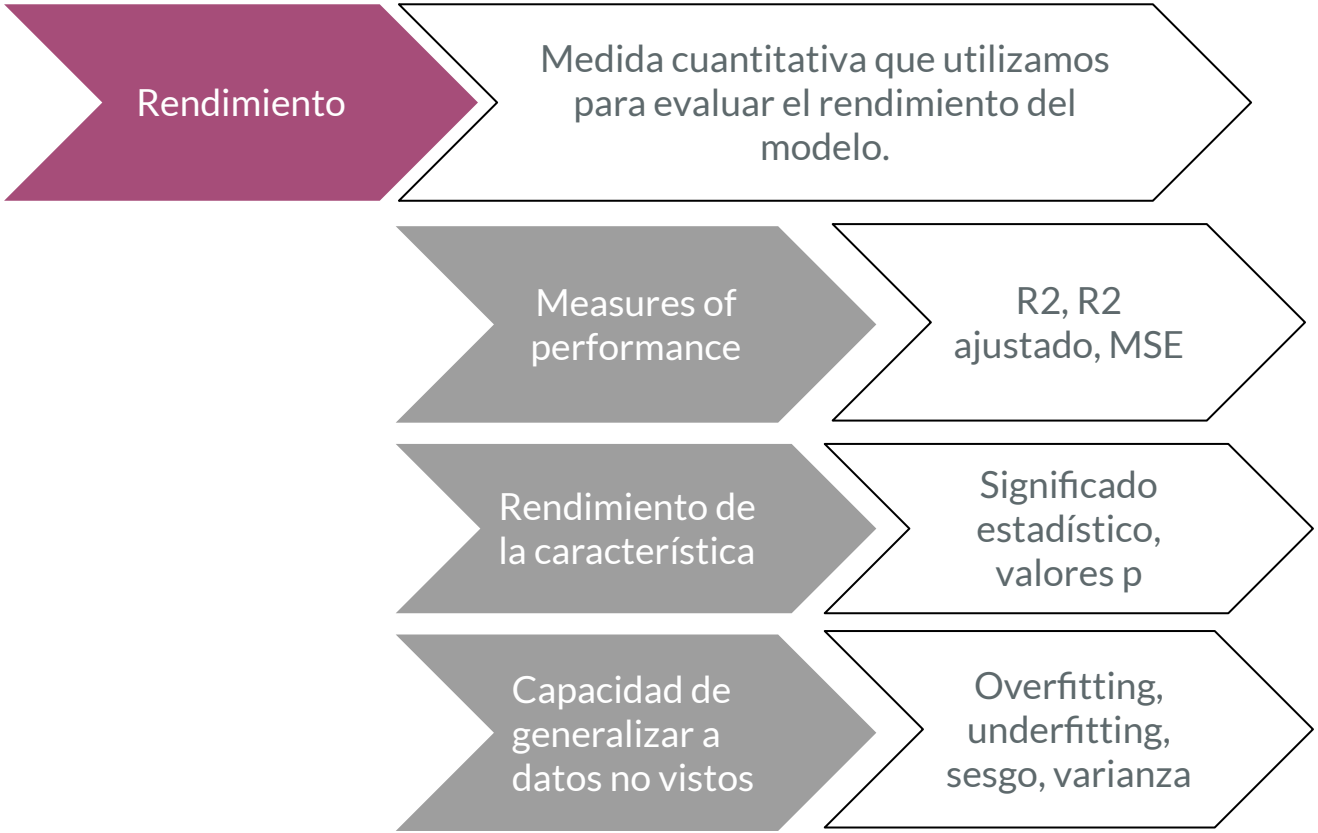
Hoy estamos mirando más de cerca cada componente del dataframe que discutimos en la última clase



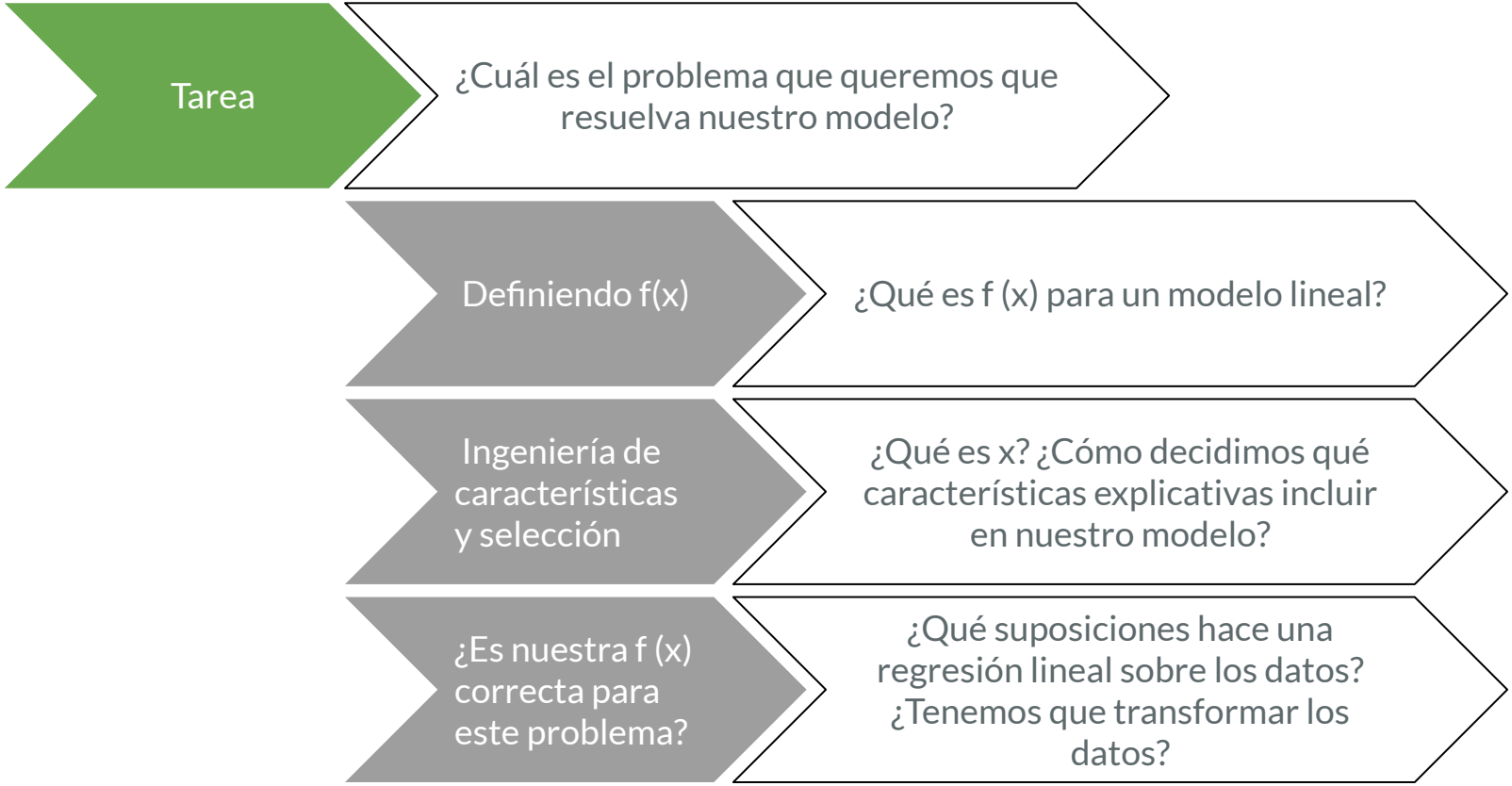
La metodología de aprendizaje es cómo el modelo de regresión lineal aprende qué línea se ajusta mejor a los datos sin procesar.



Hay modelos lineales para problemas de regresión y clasificación.



La Tarea


Tarea
¿Cuál es el problema que queremos que resuelva nuestro modelo?
Definiendo f(x)
¿Qué es f (x) para un modelo lineal?
Ingeniería de características y selección
¿Qué es x? ¿Cómo decidimos qué características explicativas incluir en nuestro modelo?
¿Es nuestra f (x) correcta para este problema?
¿Qué suposiciones hace una regresión lineal sobre los datos? ¿Tenemos que transformar los datos?

Tarea

Recordemos nuestra discusión de dos tipos de tareas supervisadas, regresión y clasificación. Hay modelos lineales para ambos. Hoy solo discutiremos la regresión.

Regresión

Variable continua

Regresión de Mínimos cuadrados ordinarios - Ordinary Least Squares (OLS)

OLS es un método de regresión lineal basado en minimizar la suma de los residuos al cuadrado

Clasificación

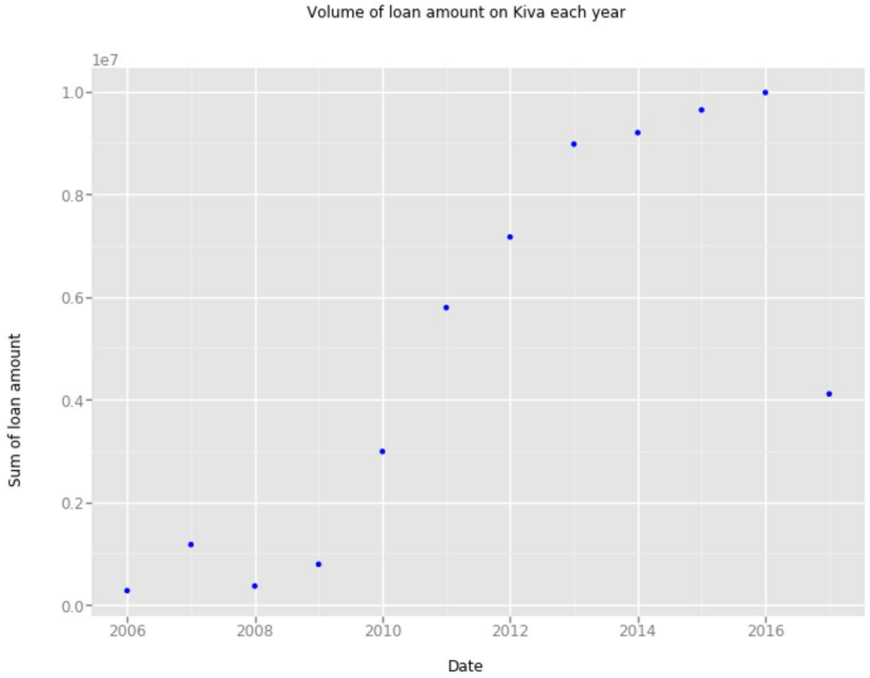
Variable categórica

Regresión logística

No cubriremos esto aquí, pero lo veremos de forma posterior.

Tarea de regresión OLS

Una regresión OLS es una línea de tendencia que predice cuánto cambiará Y para un cambio dado en x. Analicémoslo en más detalla mirando un ejemplo.



Hemos trazado el total de dinero prestado por KIVA en Kenia cada año. ¿Cuál es la tendencia? ¿Cómo resumirías esta tendencia en una oración?

Intuición humana

"Cada año, el valor de los préstamos en KE (Kenia) parece estar aumentando"

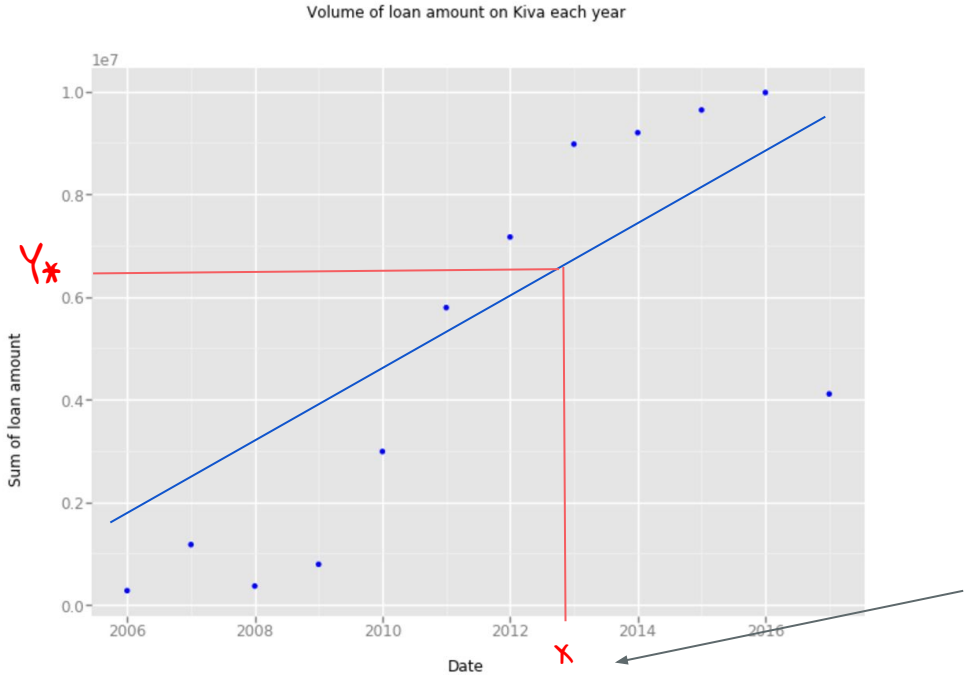
Regresión OLS

Cada año adicional corresponde a x dólares adicionales de préstamos de Kiva en KE.

Una regresión lineal formaliza nuestra tendencia percibida como una relación entre x e Y. Puede entenderse intuitivamente como una línea de tendencia.

Definiendo f(x)

Un modelo lineal expresa la relación entre nuestras características explicativas y nuestras características de resultado como una línea recta. La salida de f(x) para un modelo lineal siempre será una línea.



Un modelo lineal nos permite predecir más allá de nuestro conjunto de observaciones porque para cada punto en el eje x podemos encontrar el Y * correspondiente

Por ejemplo, ahora podemos decir para una x que no está en nuestro diagrama de dispersión (mayo de 2012), cuál es la predicción del monto de los préstamos.

¿Es la regresión lineal el modelo correcto para nuestros datos?

Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

Una gran parte de ser un investigador en este campo, implica elegir el modelo adecuado para la tarea.

Cada modelo hace ciertas suposiciones sobre los datos subyacentes.

Veamos más de cerca cómo se relaciona esto en la regresión lineal.

Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

Antes de elegir un modelo lineal, debemos asegurarnos de que todos los supuestos sean ciertos en nuestros datos.

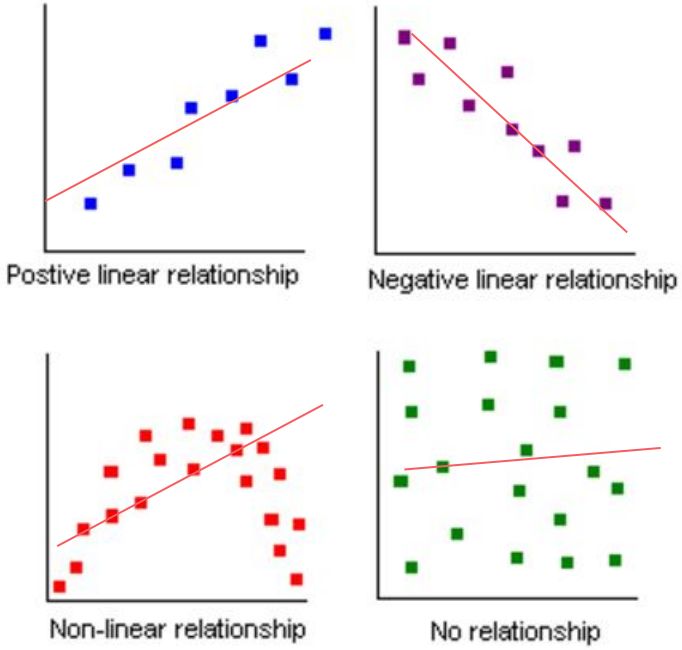
Suposiciones de la regresión lineal OLS

- ❑ Relación lineal entre x e Y
- ❑ Distribución normal de variables
- ❑ Sin autocorrelación (variables independientes)
- ❑ Homocedasticidad
- ❑ Sin multicolinealidad
- ❑ Regla general: al menos 20 observaciones por variable independiente en el análisis

Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

¿Existe una relación lineal entre x e Y?



La regresión lineal supone que existe una relación lineal entre x e Y.

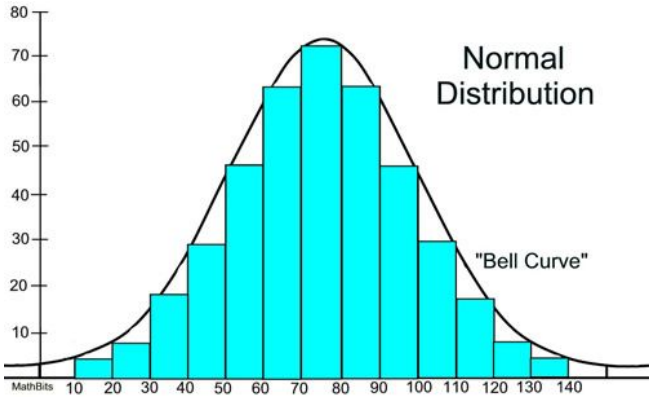
Si esto no es cierto, nuestra línea de tendencia hará un mal trabajo al predecir Y.

Fuente: Linear and non-linear relationships

Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

¿Nuestros datos se distribuyen normalmente?



La distribución normal de las características explicativas y de resultados evita la distorsión de los resultados debido a valores atípicos o datos asimétricos.

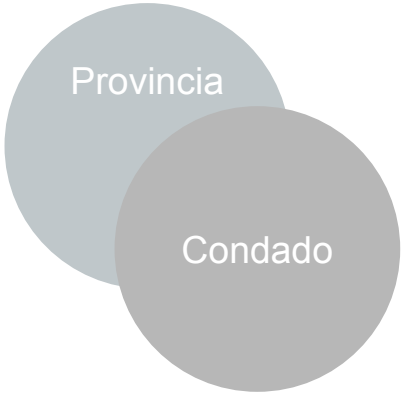
Fuente: Statistics - Normal distribution

Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

La multicolinealidad ocurre cuando las variables explicativas están altamente correlacionadas.

La multicolinealidad introduce redundancia en el modelo y reduce nuestra certeza en los resultados. No queremos multicolinealidad en nuestro modelo.



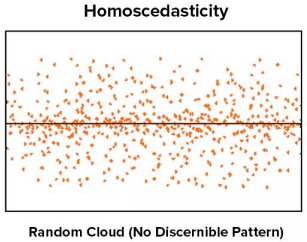
Provincia y Condado están altamente correlacionados. Queremos incluir solo una de estas variables en nuestro modelo.

La edad y el monto del préstamo parecen no tener multicolinealidad. Podemos incluir ambos en nuestro modelo.

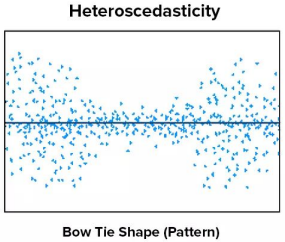
Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

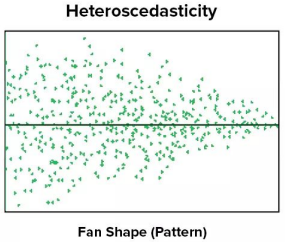
¿Tenemos homocedasticidad?



Bueno



No bueno



No bueno

Los datos deben ser homocedásticos, lo que significa que la tasa de error se distribuye uniformemente en todos los valores de las variables de resultado.

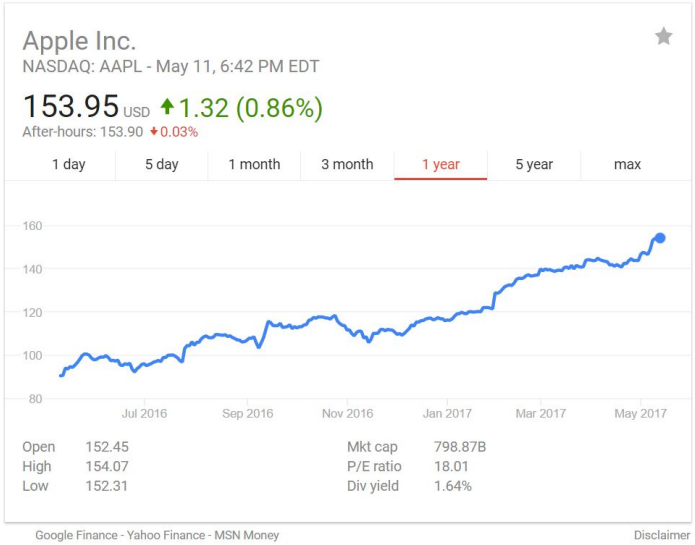
El término de error o "ruido" es el mismo en todos los valores de las variables de resultado. Si la homocedasticidad no se cumple, los casos con un mayor término de error tendrán una influencia desproporcionada en la regresión.

Fuente: Homoscedasticity

Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

¿Tenemos autocorrelación?



La autocorrelación es la correlación entre los valores de una variable y su copia retrasada. Por ejemplo, el precio de una acción hoy está correlacionado con el precio de ayer

La autocorrelación ocurre comúnmente cuando trabajas con series de tiempo.

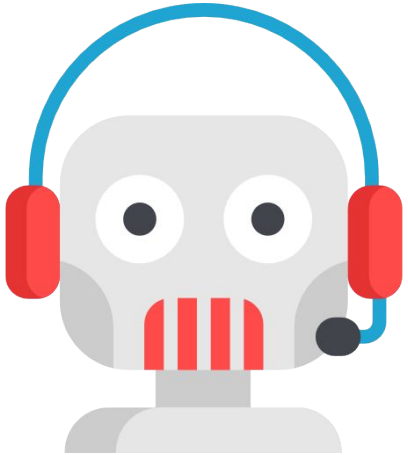
Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

En el laboratorio de codificación, revisaremos el código que te ayudará a determinar si un modelo lineal proporciona la mejor f (x)

Suposiciones OLS

- ✓ Relación lineal entre x e Y
- ✓ Distribución normal de variables
- ✓ Sin autocorrelación (variables independientes)
- ✓ Homocedasticidad
- ✓ Regla general: al menos 20 observaciones por variable independiente en el análisis



Hemos revisado todos nuestros supuestos, lo que significa que podemos elegir con confianza un modelo OLS lineal para esta tarea.

¡Comencemos a construir nuestro modelo!

Tarea de regresión OLS

¿Es nuestra f(x) correcta para este problema?

Pregunta ¿Qué sucede si no tienes todas las suposiciones?

Suposiciones OLS

- ✓ Relación lineal entre x e Y
- ✓ Distribución normal de variables
- ✓ Sin autocorrelación (variables independientes)
- ✓ Homocedasticidad
- ✓ Regla general: al menos 20 observaciones por variable independiente en el análisis

Si estos supuestos no son ciertos, nuestra línea de tendencia no será precisa. Tenemos algunas opciones:

1) **Transformar nuestros datos** para cumplir con los supuestos

2) **Elegir un modelo diferente** para capturar la relación entre x e Y

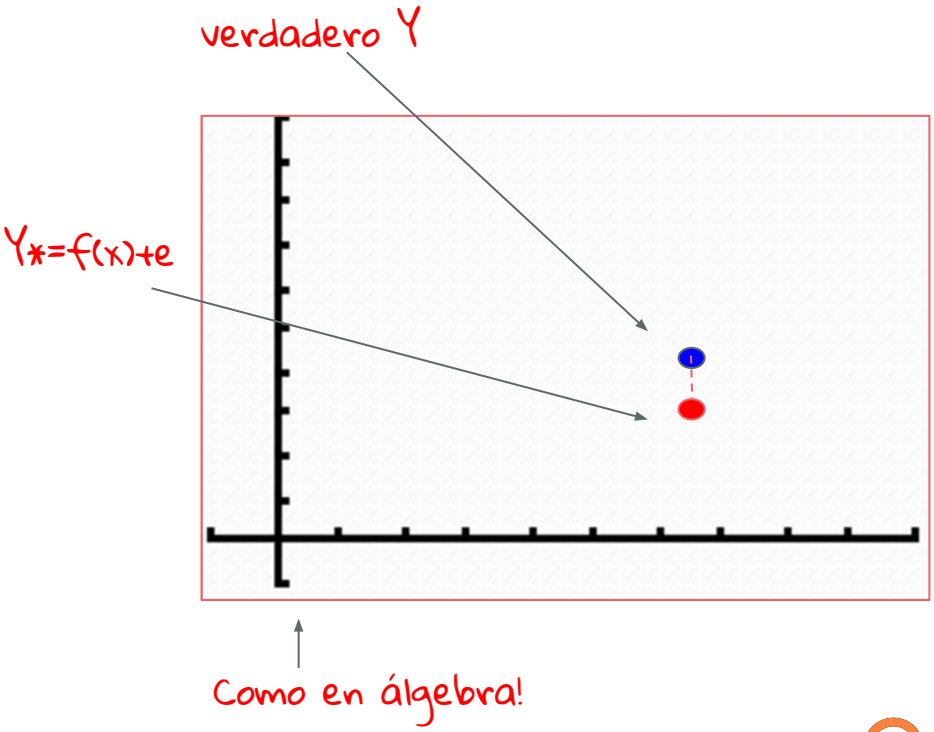
¡Sí! La regresión lineal es una opción de modelo apropiada. ¿Ahora que?

¿Qué es f(x) para un modelo de regresión lineal?

Recuerde que todos los modelos involucran una función f(x) que asigna una entrada x a una Y pronosticada (Y*). El objetivo de la función es tener un modelo que prediga Y * lo más cerca posible al Y verdadero.

f(x) para un modelo lineal es:

$$Y* = a + bx + e$$

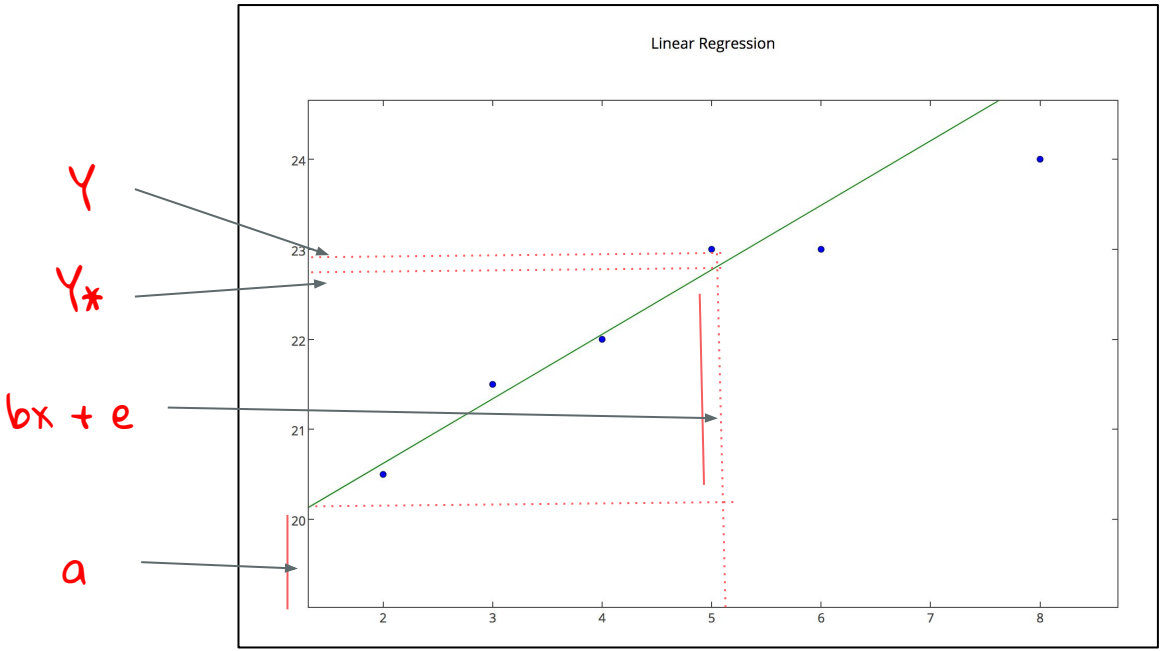


¿Qué significa realmente $Y*=a + bx + e$?

$$Y*=a + bx + e$$

parámetros

a	La intersección en y de la línea de regresión
b	La pendiente o gradiente de la línea de regresión. Esto determina cuán empinada es la línea y su direccionalidad.
e	El término de error irreducible, error que nuestro modelo no puede reducir
Y*	La salida Y pronosticada de nuestra función

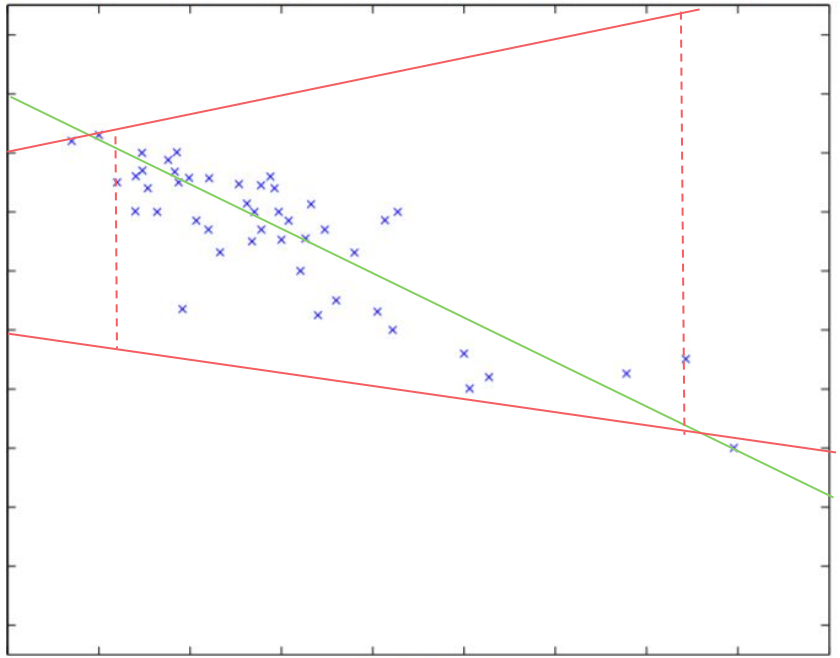


Hay un número infinito de líneas posibles en un espacio bidimensional. ¿Cómo elige nuestro modelo el mejor?

La regresión lineal es un algoritmo de aprendizaje. Eso es lo que lo convierte en un modelo de machine learning.

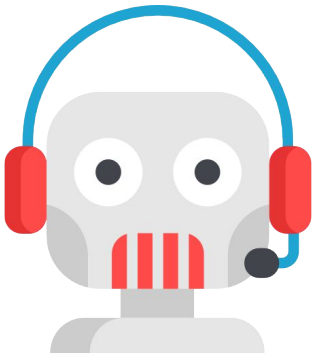
Aprende a encontrar la mejor línea de tendencia a partir de un número infinito de posibilidades. **¿Cómo?**

Primero, debemos entender qué parámetros controlamos.



Los parámetros de un modelo son valores que controlamos. Los parámetros en un modelo OLS son a (intercepción) y b (pendiente)

$$Y* = a + bx + e$$



En cada modelo hay algunas cosas que no podemos controlar, como e (error irreducible).

Un modelo también puede tener **hiperparámetros** que se configuran de antemano y no se entrenan con datos (no es necesario pensar demasiado en esto ahora).

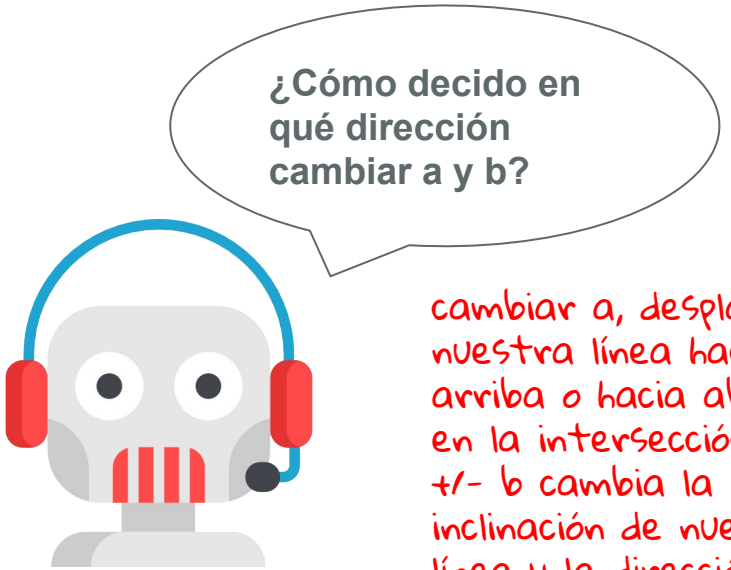
Parámetros → Valores que controlan el comportamiento del modelo y se aprenden a través de la experiencia.

Hiperparámetros → Ajustes de nivel superior de un modelo que se arreglan antes de que comience el entrenamiento.

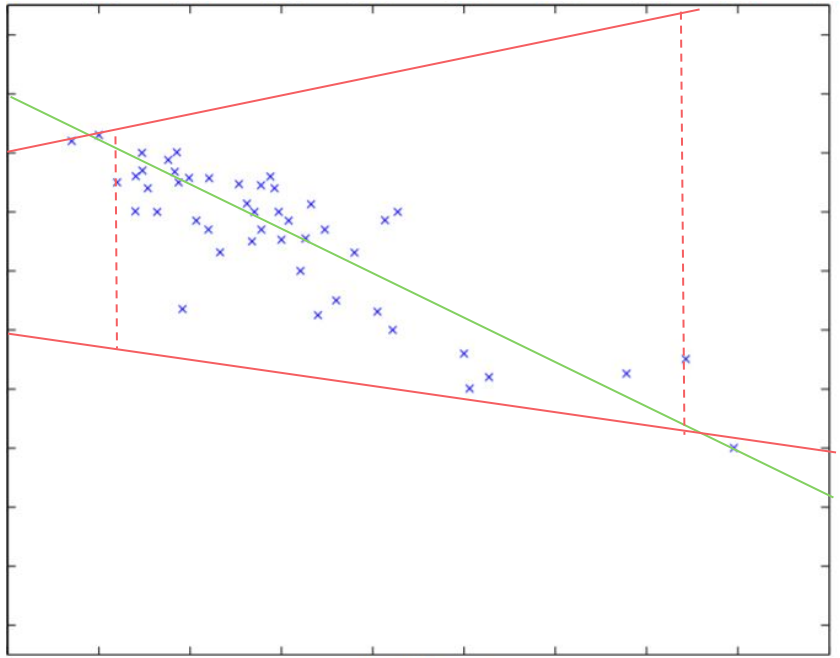
Nuestro modelo puede mover a & b pero no e.

a y b son los dos únicos parámetros que nuestro modelo OLS simple puede cambiar para acercar Y * a Y.

$$Y*=a + bx + e$$



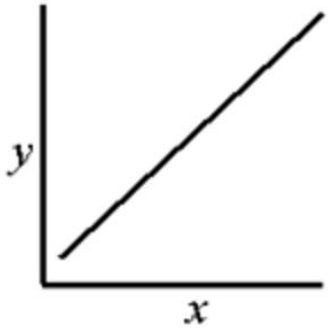
cambiar a, desplaza nuestra línea hacia arriba o hacia abajo en la intersección y, +/- b cambia la inclinación de nuestra línea y la dirección



Podemos cambiar a y b para mover nuestra línea en el espacio. A continuación hay una intuición acerca de cómo cambiar a y b afecta a f(x).

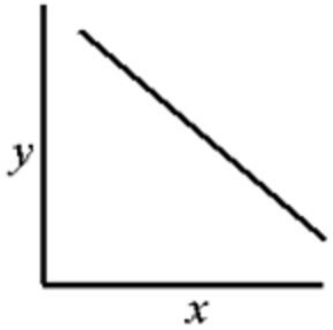
$$Y* = a + bx + e$$

a y b son nuestros dos parámetros. cambiar a, desplaza nuestra línea hacia arriba o hacia abajo en la intersección y.
a negativa mueve la intersección en y por debajo de 0.



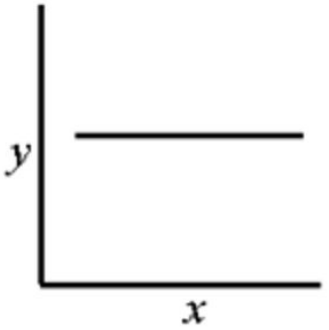
Positive slope

+ b significa una línea inclinada hacia arriba



Negative slope

- b significa una línea inclinada hacia abajo



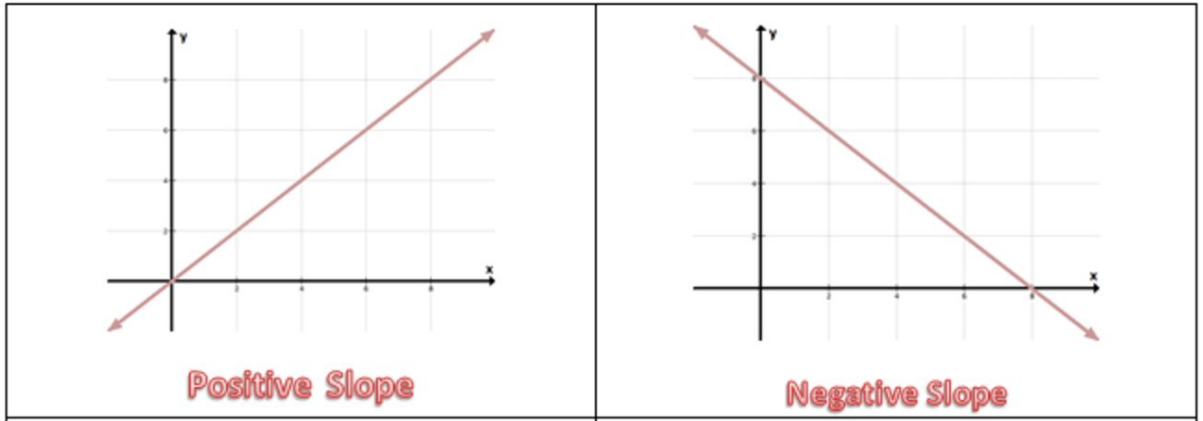
Zero slope

b = 0 significa que no hay relación entre x e Y

Fuente: Intro to Stat - Introduction to Linear Regression

La direccionalidad de b es muy importante. Nos dice si Y se hace más pequeño o más grande cuando aumentamos x.

$$Y* = a + bx + e$$



As ***x* gets larger, *y* gets larger.** **Example:** The amount of money you make (*y*) based on the number of hours you work (*x*).	As ***x* gets larger, *y* gets smaller.** **Example:** The amount of money you have left (*y*) based on the number of hours you shop (*x*).

+ b tiene una pendiente positiva

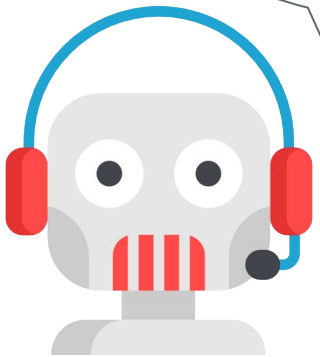
- b tiene una pendiente negativa

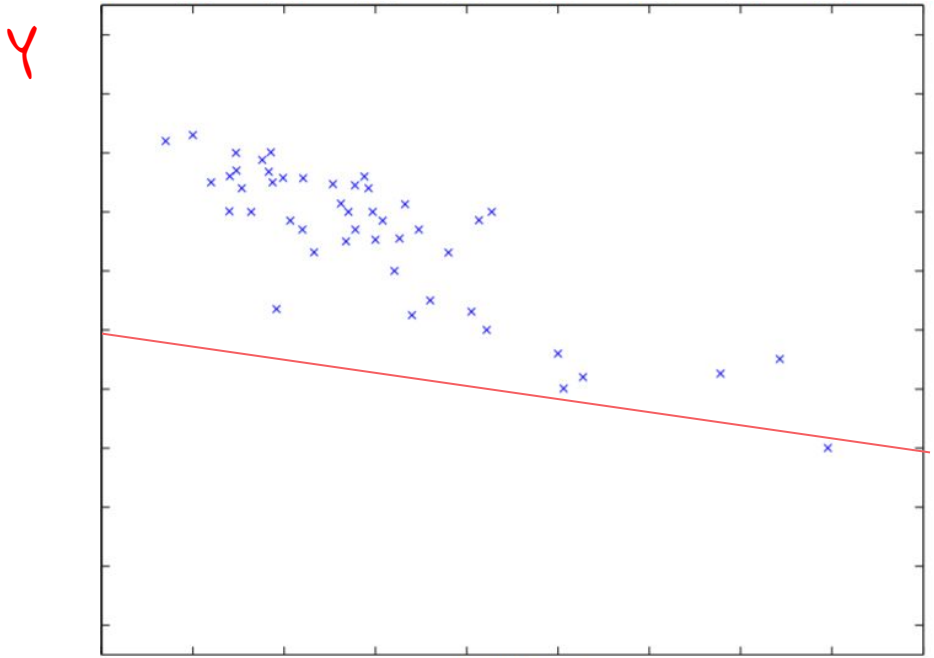
Fuente: Beginning Algebra - Coordinate System and graphing lines

Pon a prueba tu intuición.

$Y* = a + bx + e$

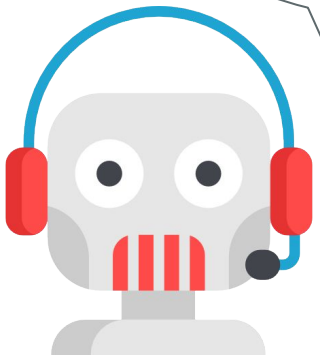
¿Qué sucede si aumento a en 2?

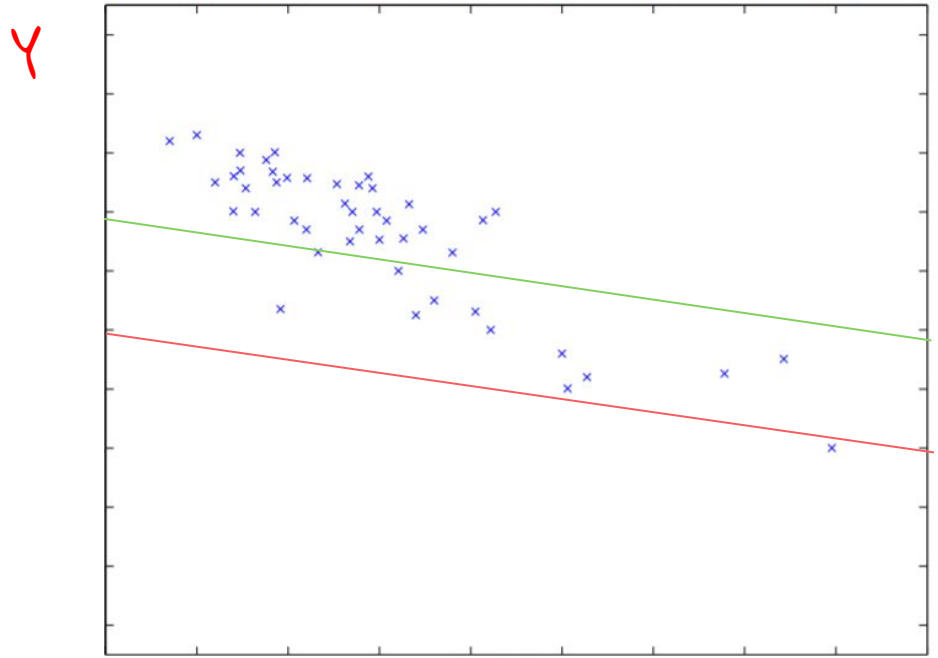




$Y* = a + bx + e$

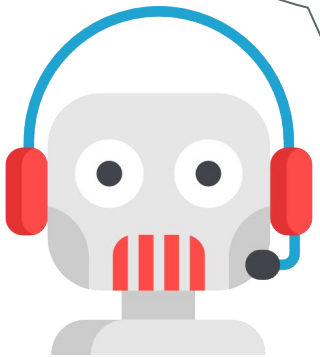
¿Qué sucede si aumento a en 2?

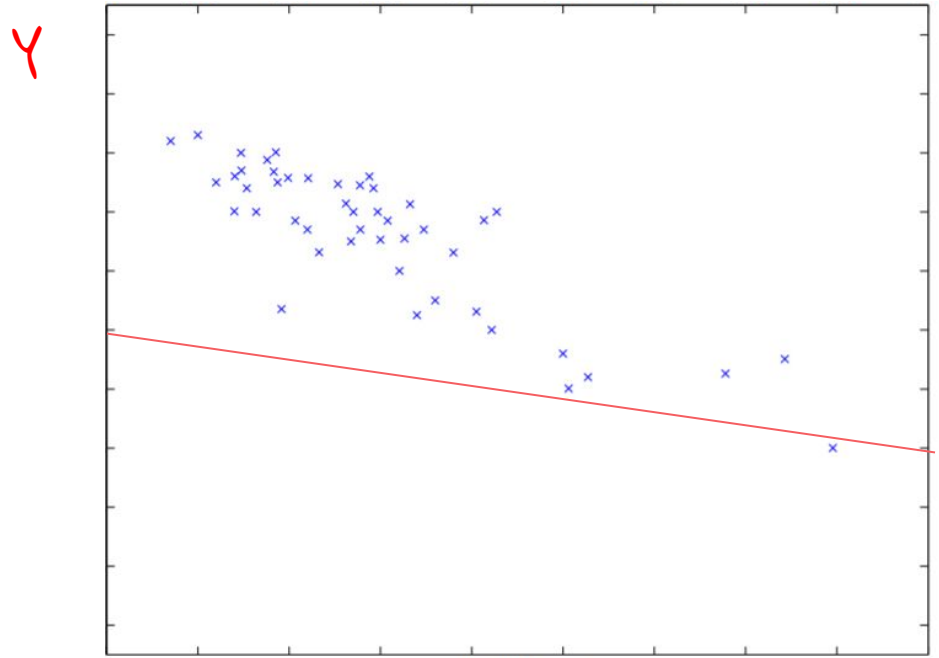




$Y* = a + bx + e$

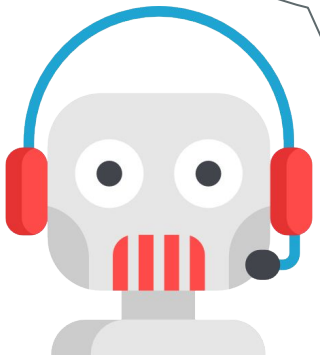
¿Qué sucede si aumento b en 3?

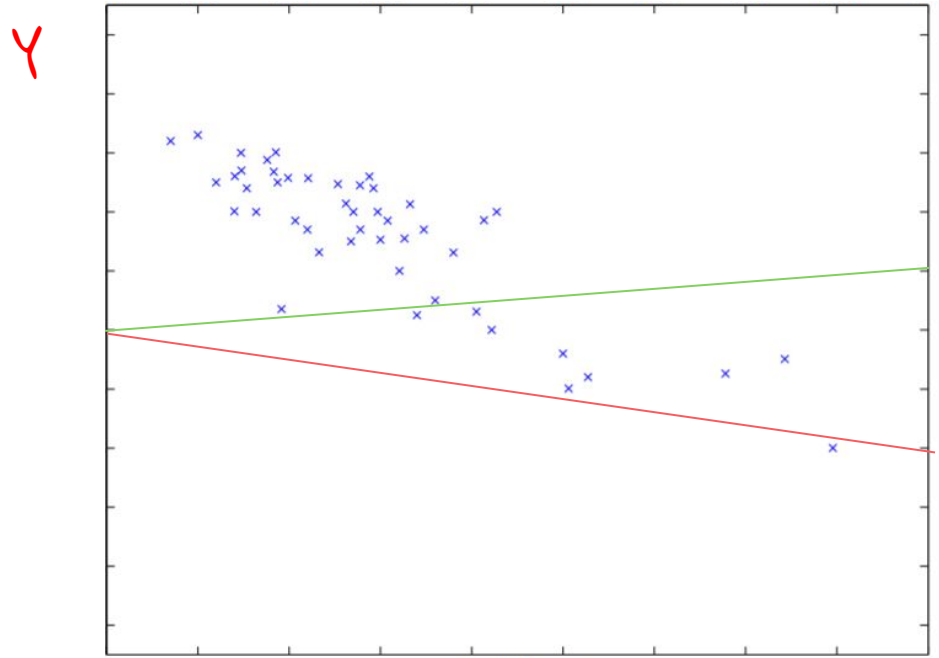




$Y* = a + bx + e$

¿Qué sucede si aumento b en 3?





¿Qué parámetros nos dan la mejor línea de tendencia?

Tomamos una decisión sobre cómo cambiar nuestros parámetros en función de nuestra función de pérdida.

$$Y* = a + bx + e$$

Recordar: **Nuestro modelo comienza con un a y un b aleatorios**, y nuestro trabajo es cambiar a y b de una manera que acerque Y* a la Y verdadera.

De hecho, estás tratando de reducir la distancia entre Y * y Y verdadero. Medimos la distancia utilizando el error cuadrático medio. Esta es nuestra función de pérdida.

Todos los modelos supervisados tienen una función de pérdida (a veces también conocida como función de costo) que deben optimizar cambiando los parámetros del modelo.

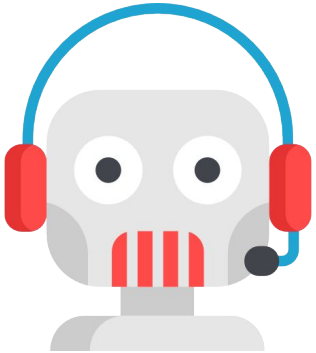
Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

Para cada a y b que probamos, medimos el error cuadrático medio. ¿Recuerdas MSE?

El error cuadrático medio es una medida de qué tan cerca está Y* de Y.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



Este trabajo no está terminado hasta que reduzca el MSE.

Hay cuatro pasos para el MSE:

Y-Y*	Para cada punto de nuestro conjunto de datos, medir la diferencia entre Y verdadero e Y pronosticado.
^2	Elevar al cuadrado cada Y-Y * para obtener la distancia absoluta, de modo que los valores positivos no cancelen los negativos cuando sumamos.
Sum	Sumar todas las observaciones para obtener el error total.
mean	Dividir la suma por el número de observaciones que tenemos.

Fuente: Intro to Stat -Introduction to Linear Regression

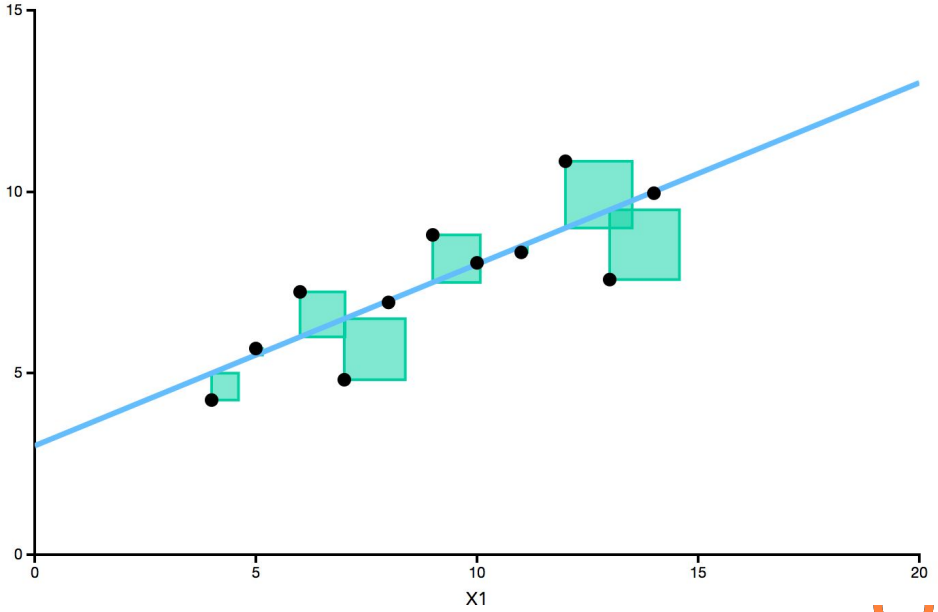
Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

Los cuadros en el gráfico a continuación son el MSE para cada punto de datos. Sumamos y luego tomamos la media en todos los puntos de datos para obtener el MSE.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Y-Y*	Para cada punto de nuestro conjunto de datos, medir la diferencia entre Y verdadero e Y pronosticado.
^2	Elevar al cuadrado cada Y-Y * para obtener la distancia absoluta, de modo que los valores positivos no cancelen los negativos cuando sumamos.
Suma	Sumar todas las observaciones para obtener el error total.
media	Dividir la suma por el número de observaciones que tenemos.



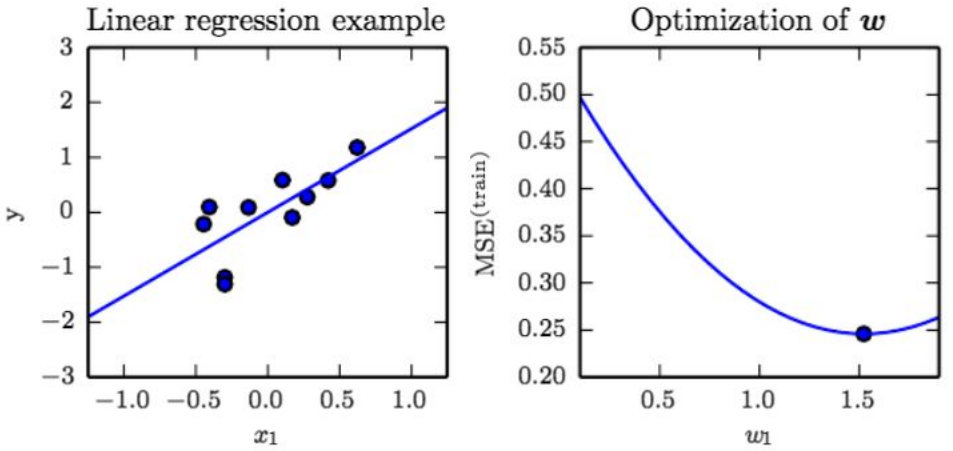
Fuente: Seeing Theory - Regression

El proceso de cambiar a y b para reducir MSE se llama **aprendizaje**. Es lo que hace que la regresión OLS sea un algoritmo de machine learning.

Para cada combinación de a y b que elegimos hay un MSE asociado. El proceso de aprendizaje implica actualizar a y b para alcanzar el mínimo global.

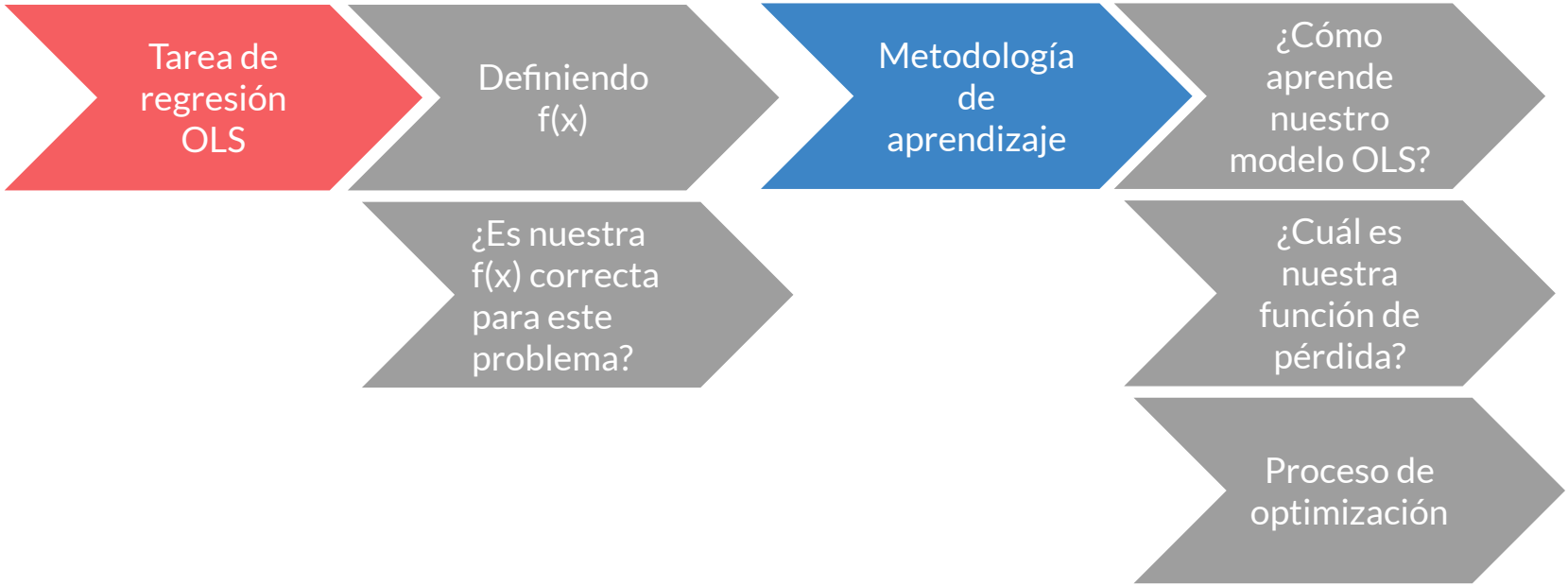
El proceso de aprendizaje para OLS se denomina técnicamente **aprendizaje por descenso de gradiente**.

Optimización de MSE:



Source: Seeing Theory - Regression

A continuación, veamos la validación del modelo.



Validación del Modelo

La validación del modelo es crucial.

La validación es un proceso de evaluación del rendimiento de su modelo.

Tenemos que evaluar un modelo en dos aspectos críticos:

1. **Qué tan cerca está la Y * del modelo de la Y verdadera**
2. **Qué tan bien funciona el modelo en el mundo real (es decir, en datos no vistos).**

Métricas comunes de validación:

1. R^2
2. R^2 ajustado
3. MSE (una función de pérdida y una medida de rendimiento)

Echemos un vistazo a estas métricas ...

Para una regresión OLS, presentaremos tres medidas de rendimiento del modelo: R^2, R^2 ajustado y MSE.

Error medio cuadrado

Ya hemos introducido MSE. Esto sirve como una función de pérdida y una evaluación del rendimiento.

R^2

Variación explicada / Variación total
Aumenta a medida que aumenta el número de x

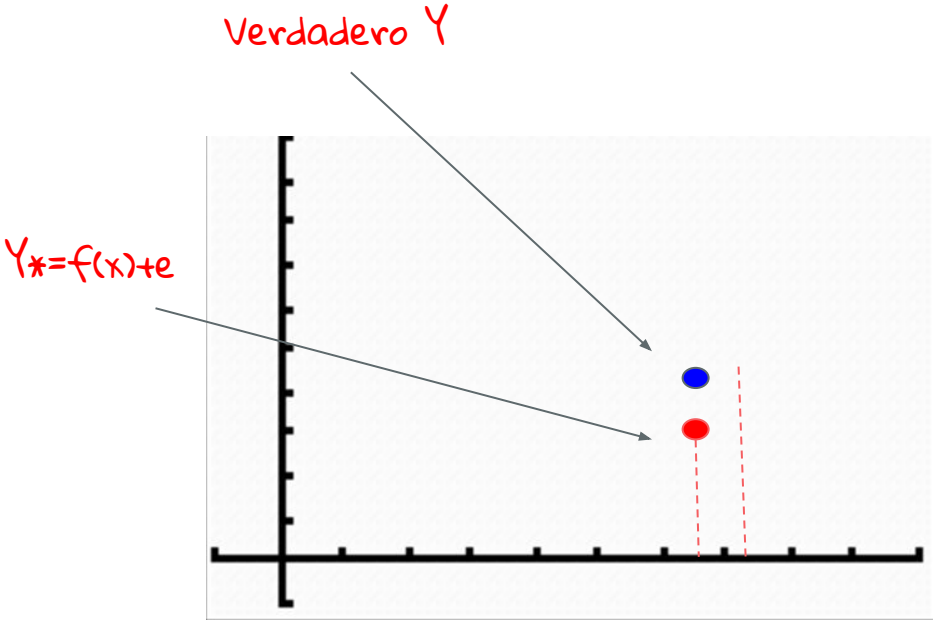
R^2 ajustado

Variación explicada / Variación total, ajustada por el número de características presentes en el modelo

R^2 y R^2 ajustado responden a la pregunta, ¿cuánto de la variación de y puede ser explicado por nuestro modelo?

Esto nos recuerda a MSE, pero las **métricas de R2 se escalan para estar entre 0 y 1.**

Nota: R^2 ajustado es preferible a R^2, porque R^2 aumenta a medida que incluyes más características en tu modelo. Esto puede inflar artificialmente R^2.



Viendo el resultado de regresión, puedes encontrar métricas R-cuadrado aquí:

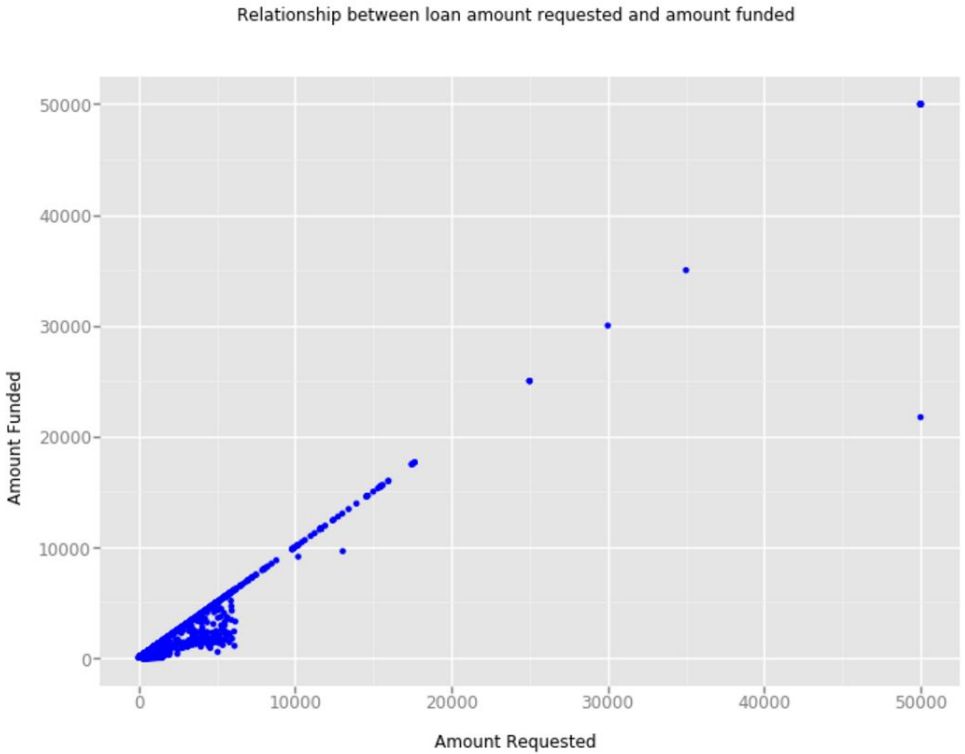
OLS Regression Results

Dep. Variable:	log_loan_amount	R-squared:	0.356
Model:	OLS	Adj. R-squared:	0.355
Method:	Least Squares	F-statistic:	1836.
Date:	Sun, 28 May 2017	Prob (F-statistic):	0.00
Time:	15:06:29	Log-Likelihood:	-73913.
No. Observations:	89811	AIC:	1.479e+05
Df Residuals:	89783	BIC:	1.481e+05
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	42.3433	1.330	31.840	0.000	39.737	44.950
sector[T.Arts]	-0.0805	0.034	-2.376	0.017	-0.147	-0.014
sector[T.Clothing]	0.0793	0.009	9.215	0.000	0.062	0.096
sector[T.Construction]	0.0292	0.017	1.736	0.083	-0.004	0.062
sector[T.Education]	-0.1096	0.015	-7.312	0.000	-0.139	-0.080

Fuente: This is a snippet of the output from Notebook 2.

R2 puede darnos causalidad donde la correlación no pudo!



R-cuadrado nos recuerda la correlación, pero hay una diferencia importante.

La correlación mide la **asociación** entre x e y.

R-cuadrado mide cuánto de **y se explica por x.**

Ahora sabemos qué tan bien funciona el modelo con los datos que tenemos, ¿cómo predecimos cómo funcionará en el mundo real?

Necesitamos una manera de cuantificar la forma en que nuestro modelo funciona con datos no vistos. En un mundo ideal, saldríamos y encontraríamos nuevos datos para probar nuestro modelo.

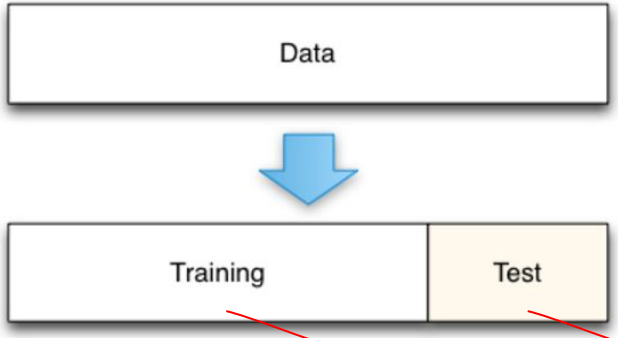
Sin embargo, esto a menudo no es realista, ya que estamos limitados por el tiempo y los recursos. En lugar de hacer esto, podemos dividir nuestros datos en dos partes: **entrenamiento y datos de prueba.**

Usaremos una parte de nuestros datos para entrenar nuestro modelo, y el resto de nuestros datos (que el modelo "no ve", como los datos del mundo real) para probar nuestro modelo.

Rendimiento

Capacidad de generalizar a datos no vistos

Dividimos nuestros datos etiquetados en entrenamiento y prueba. Los de prueba representa nuestros **datos futuros no vistos**.



Y pronosticada - Y real

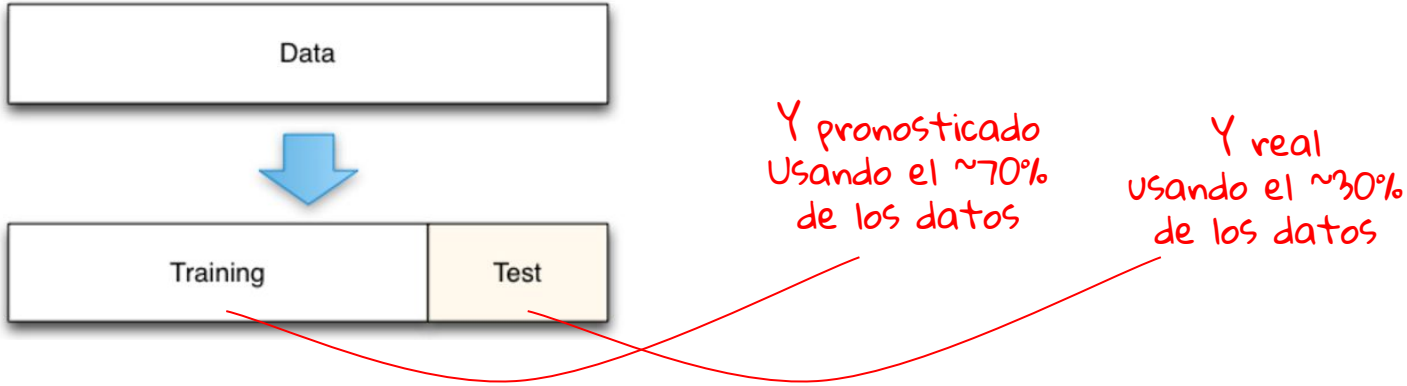
- Divide los datos al azar en conjuntos de "entrenamiento" y "prueba"
- Utilice los resultados de regresión del conjunto de "entrenamiento" para predecir el conjunto de "prueba"
- Compara Y pronosticada con Y real

Fuente: Fortmann-Roe, Accurately Measuring Model Prediction Error, http://scott.fortmann-roe.com/docs/MeasuringError.html

Rendimiento

Capacidad de generalizar a datos no vistos

Dividimos nuestros datos etiquetados en entrenamiento y prueba. Los de prueba representa nuestros **datos futuros no vistos**.



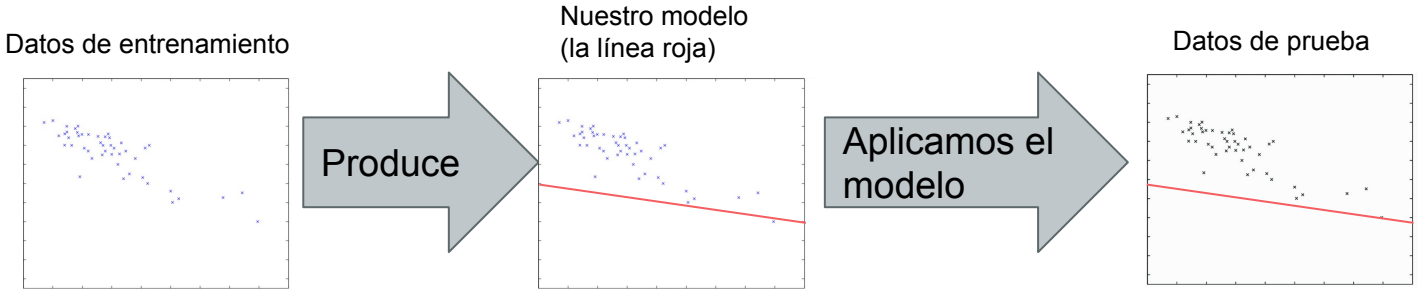
¡Los datos de prueba son "invisibles" ya que el algoritmo no los usa para entrenar el modelo!

Rendimiento

Capacidad de generalizar a datos no vistos

¡No confundir el uso de funciones de pérdida con el uso de datos de prueba!

Es importante aclarar aquí que estamos usando Y* de entrenamiento - Y entrenamiento para entrenar el modelo, **que es diferente** de Y* de prueba - Y de prueba que usamos para evaluar qué tan bien el modelo puede generalizar a datos invisibles.



Usamos Y* de entrenamiento - Y de entrenamiento en nuestras funciones de pérdida que entrenan al modelo.

Usamos Y* de prueba - Y de prueba para ver qué tan bien nuestro modelo se puede generalizar a datos invisibles, o datos que no se utilizaron para entrenar nuestro modelo.

Rendimiento

Capacidad de generalizar a datos no vistos

Dividir los datos en entrenamiento y test es **extremadamente importante**!

Evaluar si un modelo puede generalizarse o no a datos no vistos es muy importante; de hecho, poder ser capaz de predecir es a menudo el objetivo de crear un modelo.

Si no evaluamos qué tan bien un modelo puede generalizarse a datos externos, **existe el peligro de que estemos creando un modelo que sea demasiado específico para nuestro conjunto de datos** - es decir, un modelo que sea GENIAL para predecir este conjunto de datos en particular, pero NO ES ÚTIL en predecir el mundo real.

¿A qué se parece esto?

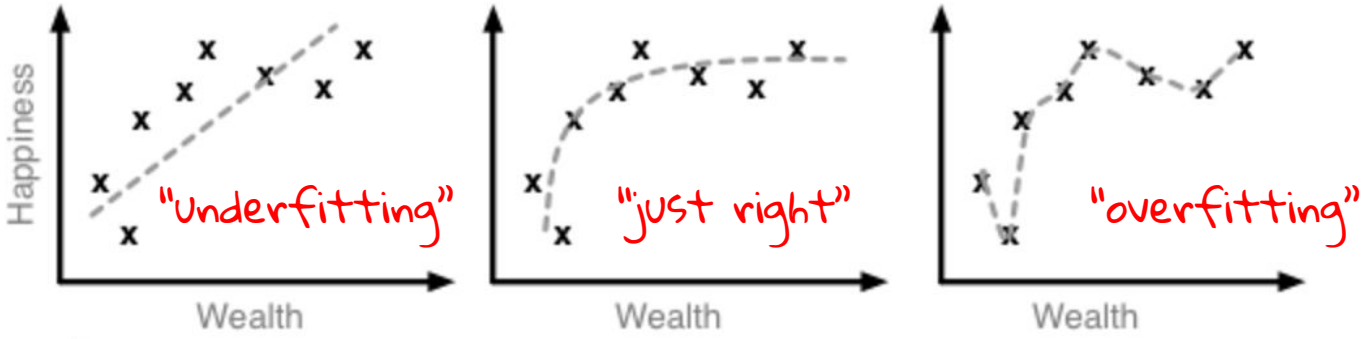
Rendimiento

Capacidad de generalizar a datos no vistos

Dividir los datos en entrenamiento y prueba es **extremadamente importante**!

Aquí, estamos usando la riqueza para predecir la felicidad. Las líneas punteadas son los modelos generados por los algoritmos de ML.

- A la izquierda, el modelo no es útil porque **es demasiado general** y no captura la relación con precisión.
- A la derecha, el modelo no es útil porque **no es lo suficientemente general** y captura todas las idiosincrasias del conjunto de datos. Esto significa que el modelo es demasiado específico para este conjunto de datos y no puede generalizarse a diferentes conjuntos de datos.



Queremos tener un modelo que sea casi perfecto: lo suficientemente preciso dentro del conjunto de datos y lo suficientemente general como para aplicarse fuera del conjunto de datos!

Rendimiento

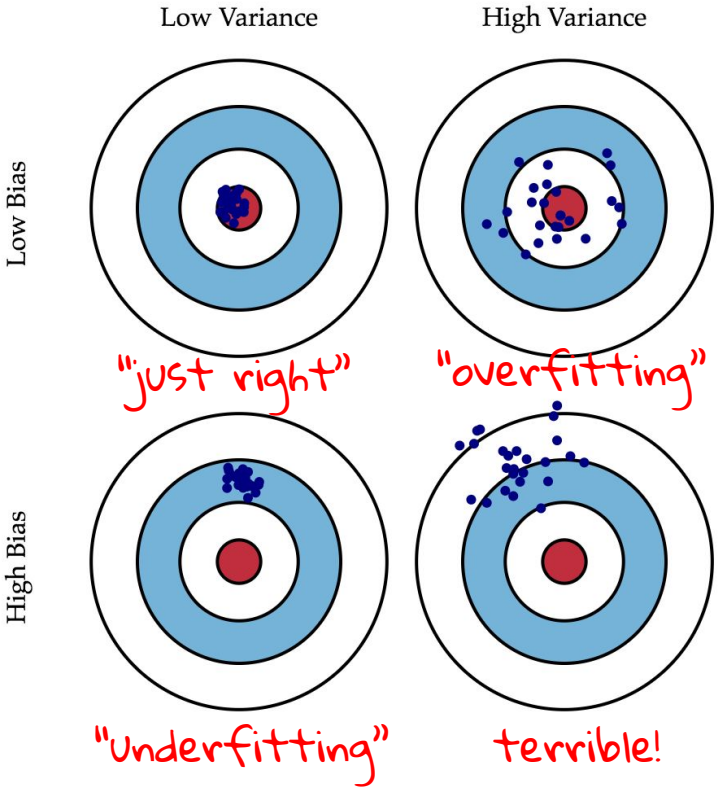
Capacidad de generalizar a datos no vistos

Este concepto de un modelo que es "just right" también se llama **término medio entre varianza y sesgo.**

El sesgo es cuán preciso es el modelo para predecir el conjunto de datos.

La varianza ies cuán sensible es el modelo a pequeñas fluctuaciones en el conjunto de entrenamiento.

Idealmente, tendríamos un sesgo bajo y una varianza baja.



Fuente: Fortmann-Roe, Accurately Measuring Model Prediction Error.
http://scott.fortmann-roe.com/docs/MeasuringError.html

No te preocupe por los detalles específicos de la compensación de la variación de sesgo por ahora: volveremos a este concepto regularmente a lo largo del curso y en profundidad en el próximo módulo.

Pasemos a otro aspecto del rendimiento que es muy importante: el rendimiento de las características.

Rendimiento

Importancia de la característica

El rendimiento de los componentes del modelo es tan importante como el rendimiento del modelo en sí.

Comprender la **importancia de las características** nos permite:

1) Cuantificar qué características estás impulsando el poder explicativo en el modelo
2) Comparar una característica con otra
3) Guiar a la selección final de características

Evaluar el rendimiento de las características se vuelve importante cuando comenzamos a usar modelos lineales más sofisticados donde f(x) incluye más de una variable explicativa. Veamos ese concepto y luego regresemos aquí.

Tarea OLS

Univariante vs.Multivariante como f (x)1

Decidir si una regresión univariada o multivariada es el mejor modelo para tu problema es parte de la tarea.

Univariante

Una variable explicativa

P.ej. Tratando de predecir la malaria usando solo la temperatura del paciente

Multivariante

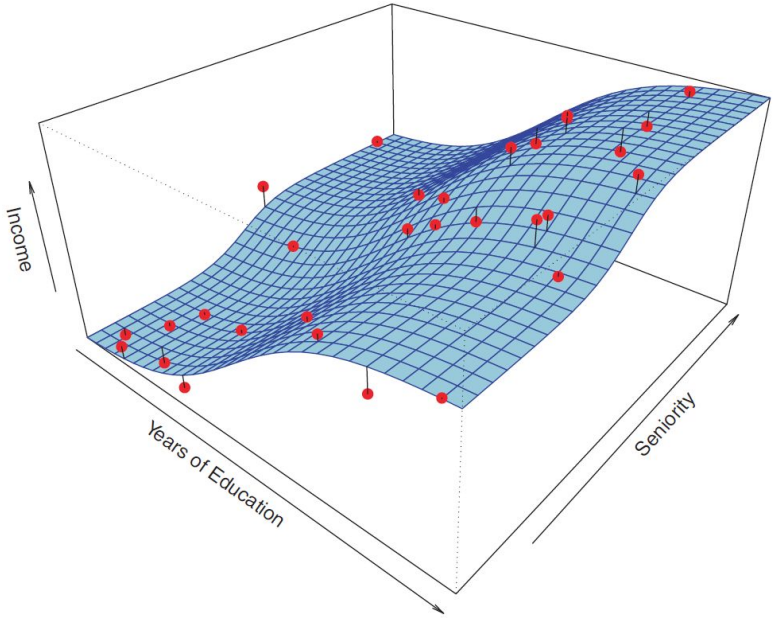
Múltiples variables explicativas.

P.ej. Intentar predecir la malaria usando la temperatura del paciente, el historial de viaje, si tienen escalofríos, náuseas o dolor de cabeza.

Tarea OLS

Definiendo f(x)

Muchas de las relaciones que probamos y modelamos son más complicadas que una regresión univariada. En cambio, usamos una regresión multivariada.

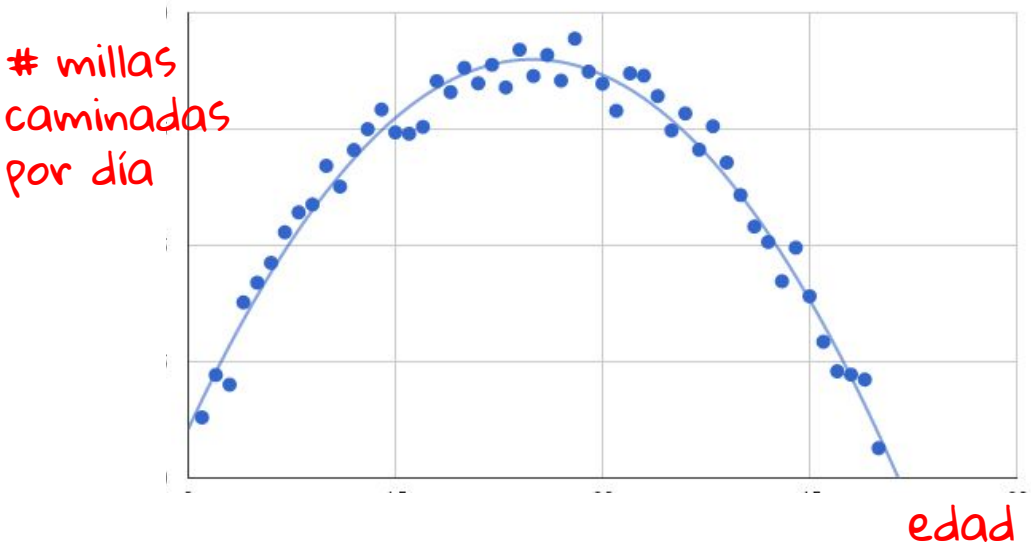


Ingresos = $a + b_1(\text{Antigüedad}) + b_2(\text{Years of Educación}) + e$

Este es un ejemplo de regresión lineal con 2 variables explicativas en 3 dimensiones. **Extender esto a n variables en n + 1 dimensiones.**

Fuente: James, Witten, Hastie and Tibshirani. Introduction to Statistical Learning with Applications in R.

¡Las regresiones también pueden ser no lineales!



Millas caminadas por día =
$a + b_1(Age)^2 + b_2(Age) + e$

En este ejemplo, vemos que la relación entre tu movilidad y tu edad aumenta y luego disminuye después de algún punto. Este es un **ejemplo de una regresión no lineal**, que se explica mejor por una ecuación cuadrática.

Fuente: [link]

OLS Task

Ingeniería de características y selección

Cuando modelamos usando una regresión multivariada, la selección de características se convierte en un paso importante. ¿Qué variables explicativas debemos incluir?

La selección de características suele ser la diferencia entre un proyecto que falla y tiene éxito. Algunas formas comunes de hacer la selección de funciones:

Investigación cualitativa

Revisión bibliográfica de trabajos anteriores realizados

Usando otros modelos

Cuantifica la importancia de la característica (veremos esto más adelante con árboles de decisión)

Análisis exploratorio

Controles de sanidad en los datos; intuición humana de lo que influiría en el resultado

Análisis de la salida de la regresión lineal

Mira el coeficiente y el valor-p de cada característica

Rendimiento

Importancia de la característica

¿Cómo evaluamos la importancia de las características en una regresión lineal?

OLS Regression Results

Dep. Variable:	log_loan_amount	R-squared:	0.356
Model:	OLS	Adj. R-squared:	0.355
Method:	Least Squares	F-statistic:	1836.
Date:	Sun, 28 May 2017	Prob (F-statistic):	0.00
Time:	15:06:29	Log-Likelihood:	-73913.
No. Observations:			
Df Residuals:			
Df Model:			
Covariance Type:	nonrobust		

Cada característica tiene un **coeficiente** y un **valor-p**.

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	42.3433	1.330	31.840	0.000	39.737	44.950
sector[T.Arts]	-0.0805	0.034	-2.376	0.017	-0.147	-0.014
sector[T.Clothing]	0.0793	0.009	9.215	0.000	0.062	0.096
sector[T.Construction]	0.0292	0.017	1.736	0.083	-0.004	0.062
sector[T.Education]	-0.1096	0.015	-7.312	0.000	-0.139	-0.080

¿Cómo evaluamos la importancia de las características en una regresión lineal?

Cada característica tiene un:

1. **Coeficiente**
2. **Valor-p**

El resultado de una regresión lineal es un modelo:

$$Y* = intercept + coef*feature$$

El tamaño del coeficiente = c**antidad de influencia que la característica tiene sobre y.** Si la característica es negativa o positiva es la **dirección de la relación que la característica tiene con y**.

¿Cómo evaluamos la importancia de las características en una regresión lineal?

Cada característica tiene un:

1. **Coeficiente**
2. **Valor-p** Expresado como un %

Un coeficiente enorme es excelente, pero la **confianza** que tengamos en ese coeficiente **depende del valor-p de ese coeficiente.**

En términos técnicos, el valor-p es la probabilidad de obtener resultados tan extremos como los observados, si el coeficiente fuera realmente cero. Responde a la pregunta: "**¿Podría haber obtenido mi resultado por casualidad?**"

Un valor p pequeño (<= 0.05, o 5%) dice que el resultado probablemente no sea una posibilidad aleatoria, *¡una gran noticia para nuestro modelo!*

¿Cara o cruz? ¡Es al azar!

Rendimiento

Importancia de la característica

¿Cómo evaluarías esta característica utilizando el valor-p y el tamaño del coeficiente?

OLS Regression Results

Dep. Variable:	log_loan_amount	R-squared:	0.356
Model:	OLS	Adj. R-squared:	0.355
Method:	Least Squares	F-statistic:	1836.
Date:	Sun, 28 May 2017	Prob (F-statistic):	0.00
Time:	15:06:29	Log-Likelihood:	-73913.
No. Observations:	89811	AIC:	1.479e+05
Df Residuals:	89783	BIC:	1.481e+05
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	42.3433	1.330	31.840	0.000	39.737	44.950
sector[T.Arts]	-0.0805	0.034	-2.376	0.017	-0.147	-0.014
sector[T.Clothing]	0.0793	0.009	9.215	0.000	0.062	0.096
sector[T.Construction]	0.0292	0.017	1.736	0.083	-0.004	0.062
sector[T.Education]	-0.1096	0.015	-7.312	0.000	-0.139	-0.080

Rendimiento

Importancia de la característica

¿Cómo evaluarías esta característica utilizando el valor-p y el tamaño del coeficiente?

OLS Regression Results

Dep. Variable:	log_loan_amount	R-squared:	0.356
Model:	OLS	Adj. R-squared:	0.355
Method:	Least Squares	F-statistic:	1836.
Date:	Sun, 28 May 2017		
Time:	15:06:29		
No. Observations:	89811		
Df Residuals:	89783		
Df Model:	27		
Covariance Type:	nonrobust		

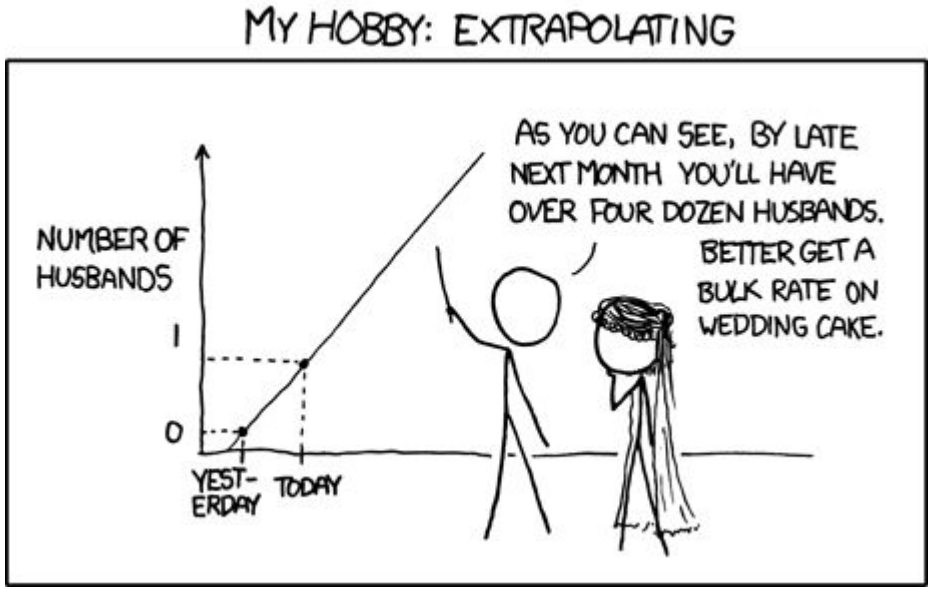
Una persona en el sector de la confección obtendrá, en promedio, un monto de préstamo más alto, pero solo por muy poco. Estoy razonablemente confiado en esta conclusión.

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	42.3433	1.330	31.840	0.000	39.737	44.950
sector[T.Arts]	-0.0805	0.034	-2.376	0.017	-0.147	-0.014
sector[T.Clothing]	0.0793	0.009	9.215	0.000	0.062	0.096
sector[T.Construction]	0.0292	0.017	1.736	0.083	-0.004	0.062
sector[T.Education]	-0.1096	0.015	-7.312	0.000	-0.139	-0.080

Unos últimos pensamientos ...

La extrapolación es el acto de inferir valores desconocidos basados en datos conocidos.

Incluso los algoritmos validados están sujetos a una extrapolación irresponsable!



La regresión lineal tiene casi innumerables aplicaciones potenciales, siempre que la interpretemos con cuidado.

Fuente: https://xkcd.com/605/

"Todos los modelos están equivocados, algunos son útiles".

- George E.P. Box,
Estadístico británico

Módulo Checklist

- ✓ Regresión lineal
  - ✓ Relación entre dos variables (x e y)
    - ✓ Formalizando f(x)
    - ✓ Correlación entre dos variables.
    - ✓ Supuestos
  - ✓ Ingeniería de características y selección
  - ✓ Regresión Univariante, regresión Multivariada
  - ✓ Medidas de rendimiento (R2, R2 ajustado, MSE)
  - ✓ Overfitting, Underfitting
  - ✓ Proceso de aprendizaje: función de pérdida y error cuadrático medio

Recursos avanzados

¿Quieres llevar esto más lejos? Aquí hay algunos recursos que recomendamos:

- Libros
  - An Introduction to Statistical Learning with Applications in R (James, Witten, Hastie and Tibshirani): Chapters 2.1, 3, 4, 6
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Hastie, Tibshirani, Friedman): Chapters 3, 4
- Recursos en línea
  - Analytics Vidhya's guide to understanding regression: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/
  - Brown University's introduction to probability and statistics, http://students.brown.edu/seeing-theory/
- Si estás interesado en modelos de regresión más sofisticados, busca:
  - Logistic regression, Polynomial regression, Interactions
- Si estás interesado en formas adicionales de resolver la multicolinealidad, busca:
  - Eigenvectors, Principal Components Analysis

Felicidades! ¡Terminaste el módulo 3!

Obtén más información sobre ML de Delta para una buena misión aquí.