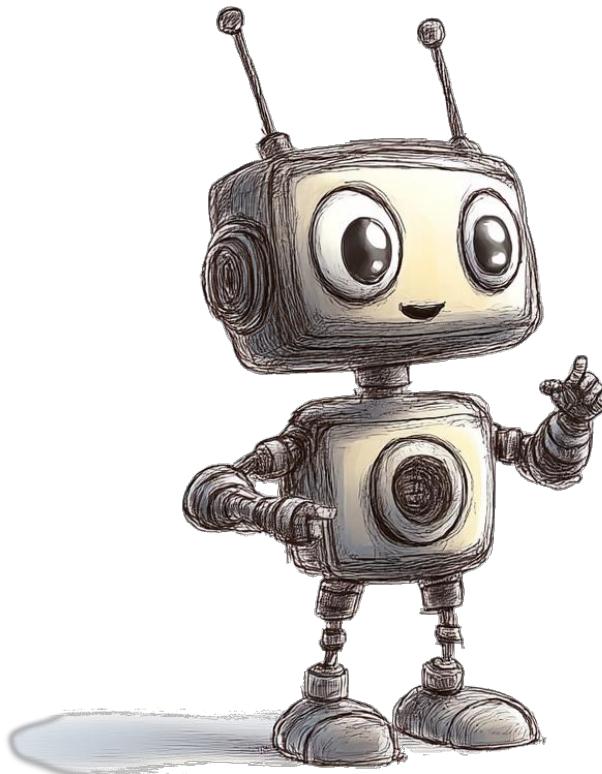
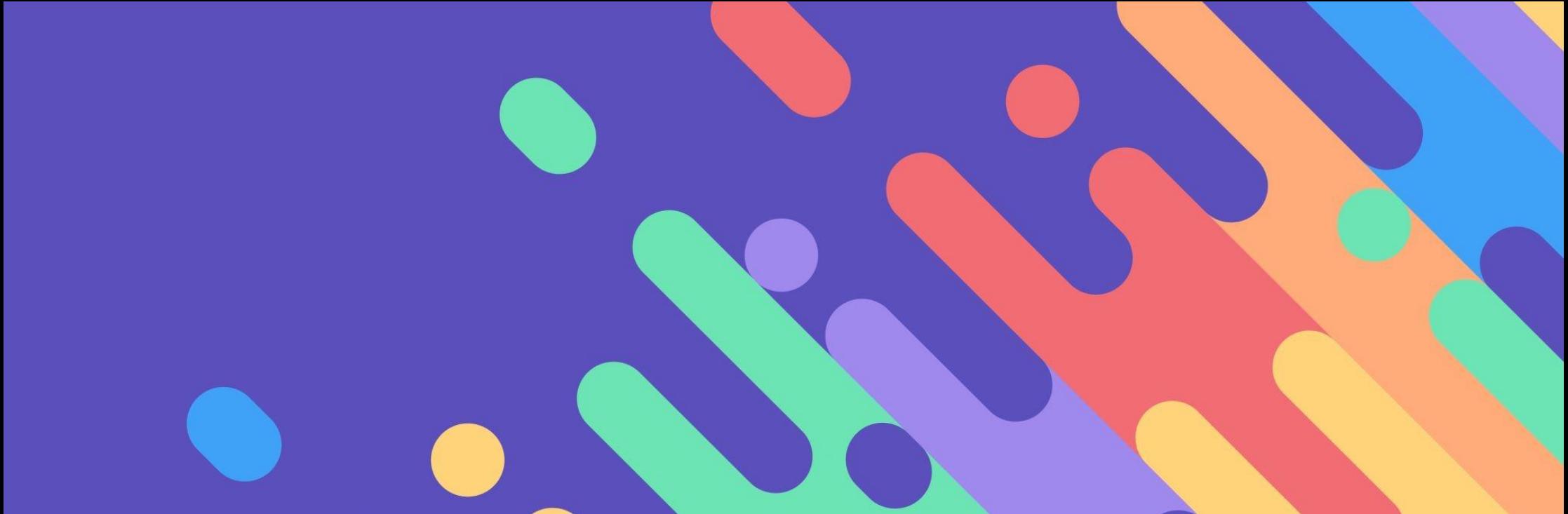


REGRESIÓN



Los temas que vamos a ver en este video son:

- Regresión Lineal
 - Definición
 - Ajuste
 - Supuestos
- Tratamiento de variables
 - Normalización
 - Estandarización
 - Variables Categóricas: One-Hot Encoding
- Métricas de evaluación

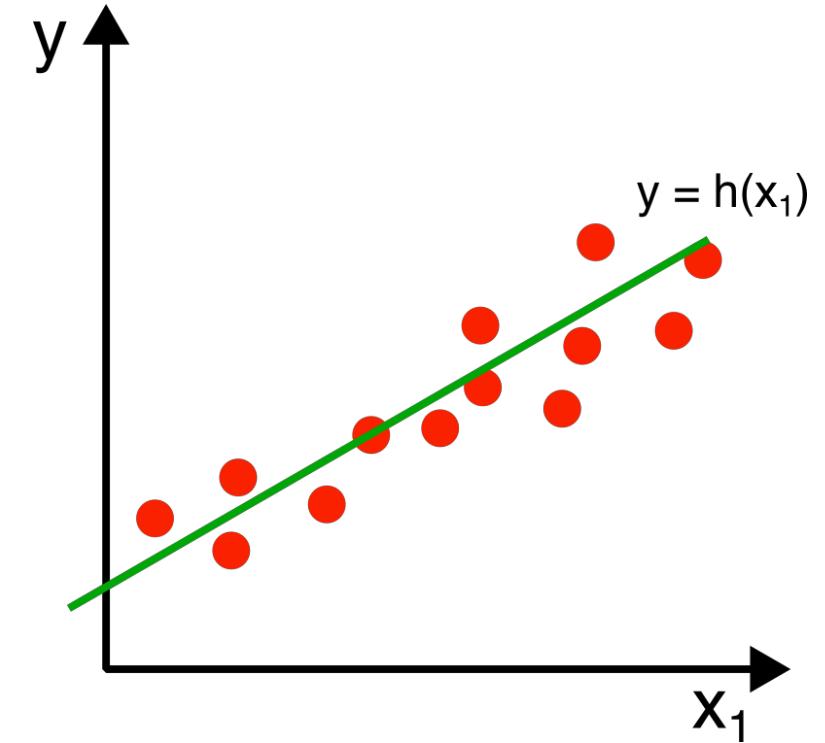


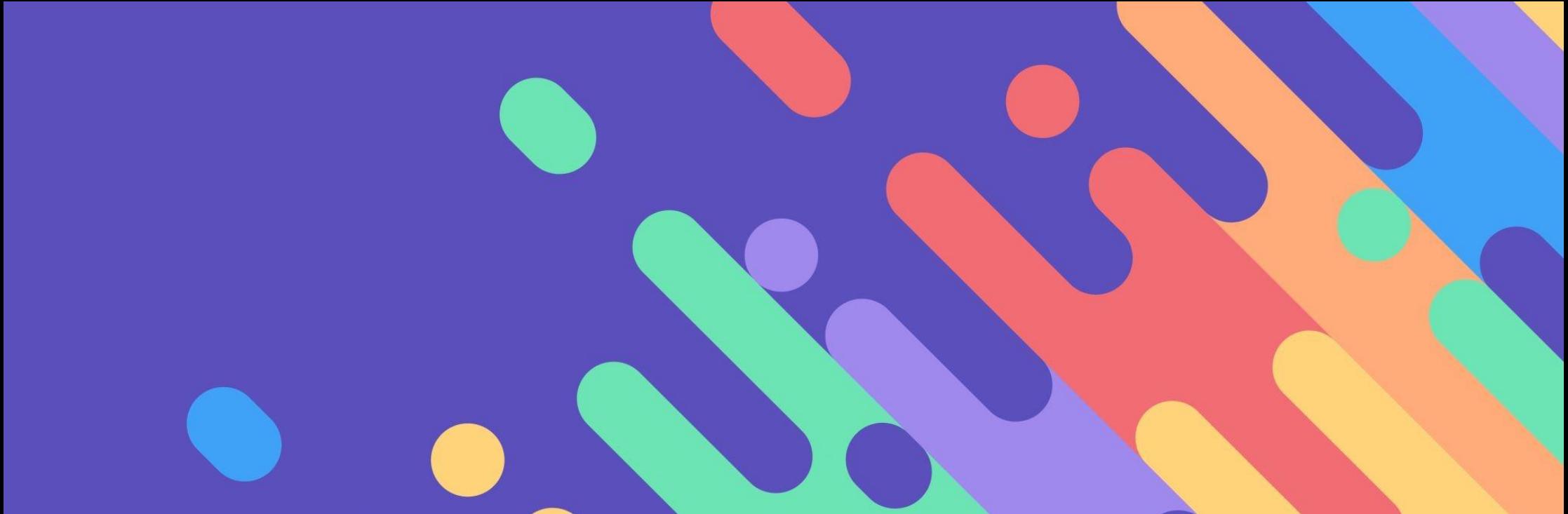
REGRESIÓN

REGRESIÓN

Si tenemos un problema donde el target **y** es una *variable numérica*, se llama un **problema de regresión**.

Se centra en estudiar las relaciones entre una variable dependiente y una o más variables independientes.





REGRESIÓN LINEAL

REGRESIÓN LINEAL

La regresión lineal es parte de una subárea del aprendizaje automático llamada **aprendizaje estadístico**, que:

- Se enfoca en el estudio de modelos estadísticos y métodos para analizar y comprender los datos.
- Utiliza herramientas y técnicas estadísticas clásicas para hacer inferencias y tomar decisiones basadas en los datos.
- Su enfoque principal puede ser la estimación de parámetros, la predicción o la inferencia sobre la relación entre variables.

REGRESIÓN LINEAL

El modelo de regresión lineal es una combinación lineal de las variables de entrada:

$$\hat{y} = h(X) = b + w_0x_0 + \cdots + w_dx_d$$

- $X = (x_0, x_1, \dots, x_d)$
- b, w_0, \dots, w_d
- \hat{y}

REGRESIÓN LINEAL

El modelo de regresión lineal es una combinación lineal de las variables de entrada:

$$\hat{y} = h(X) = b + w_0x_0 + \cdots + w_dx_d$$

- $X = (x_0, x_1, \dots, x_d)$ Son las *características (features)* de nuestras observaciones. Son todas **variables numéricas**.
- b, w_0, \dots, w_d
- \hat{y}

REGRESIÓN LINEAL

El modelo de regresión lineal es una combinación lineal de las variables de entrada:

$$\hat{y} = h(X) = b + w_0x_0 + \cdots + w_dx_d$$

- $X = (x_0, x_1, \dots, x_d)$ Son las *características (features)* de nuestras observaciones. Son todas **variables numéricas**.
- b, w_0, \dots, w_d Son los coeficientes del modelo. Son números reales. Cuanto más cerca de cero esté un coeficiente, menos depende la variable dependiente del *feature* que multiplica.
- \hat{y}

REGRESIÓN LINEAL

El modelo de regresión lineal es una combinación lineal de las variables de entrada:

$$\hat{y} = h(X) = b + w_0x_0 + \cdots + w_dx_d$$

- $X = (x_0, x_1, \dots, x_d)$ Son las *características (features)* de nuestras observaciones. Son todas **variables numéricas**.
- b, w_0, \dots, w_d Son los coeficientes del modelo. Son números reales. Cuanto más cerca de cero esté un coeficiente, menos depende la variable dependiente del *feature* que multiplica.
- \hat{y} Es la predicción del modelo. Se compara con la *etiqueta (label)* de la observación.

REGRESIÓN LINEAL

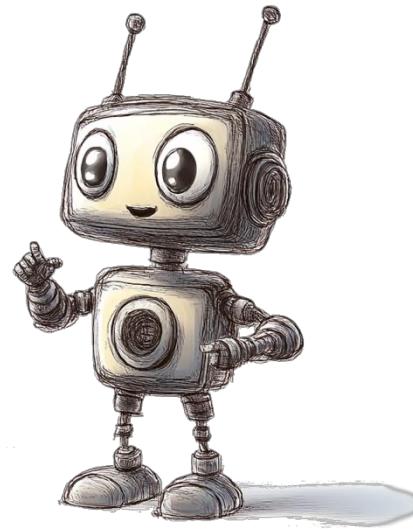
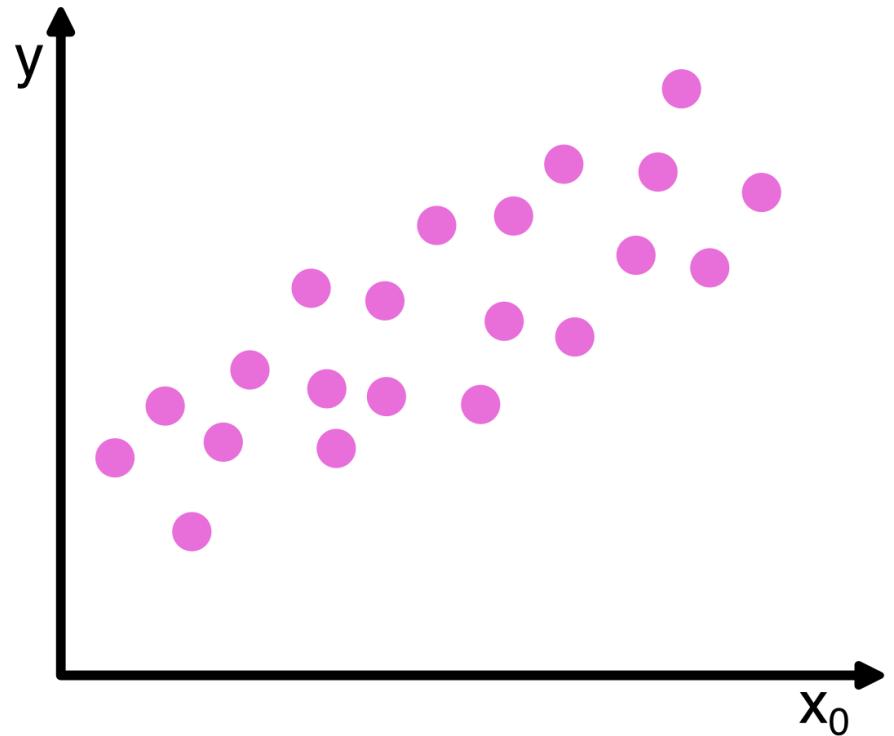
El modelo de regresión lineal es una combinación lineal de las variables de entrada:

$$\hat{y} = h(X) = b + w_0x_0 + \cdots + w_dx_d$$

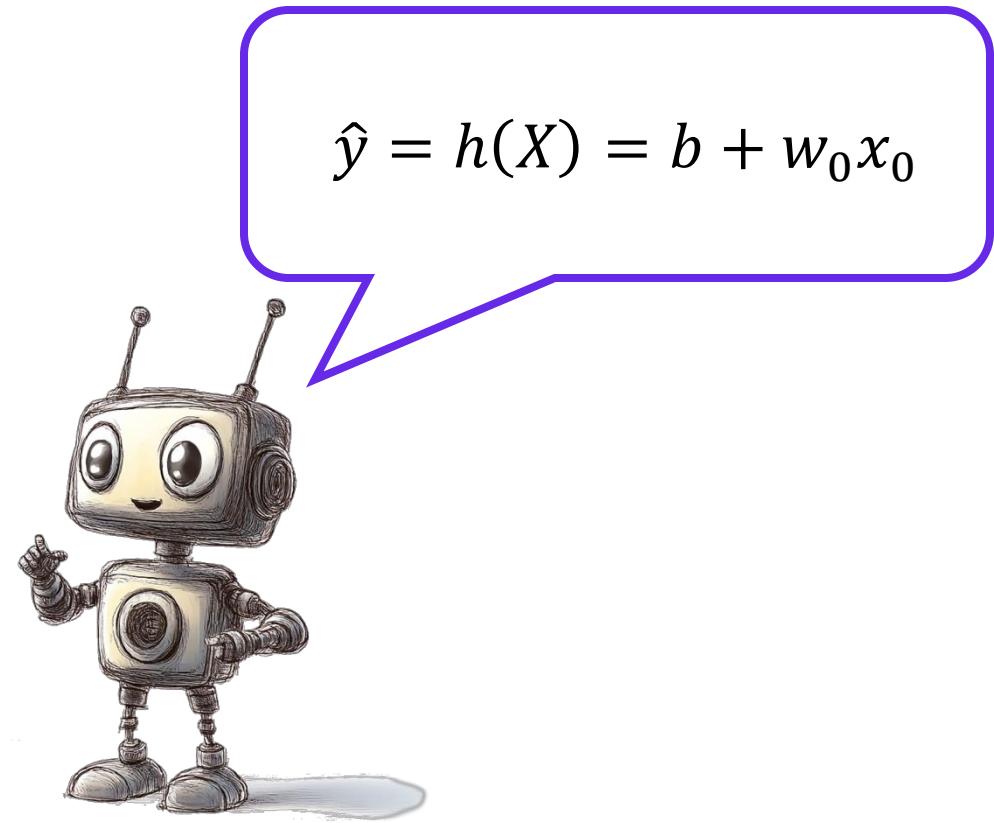
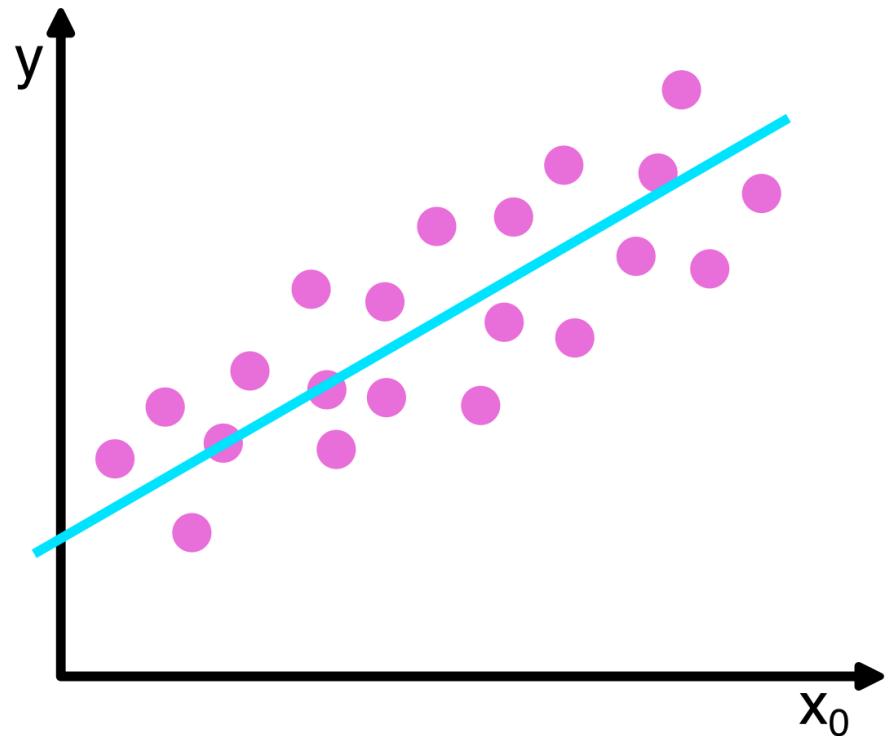
Podemos también expresarlo en notación matricial:

$$\hat{y} = h(X) = b + \mathbf{W}^T \mathbf{X}$$

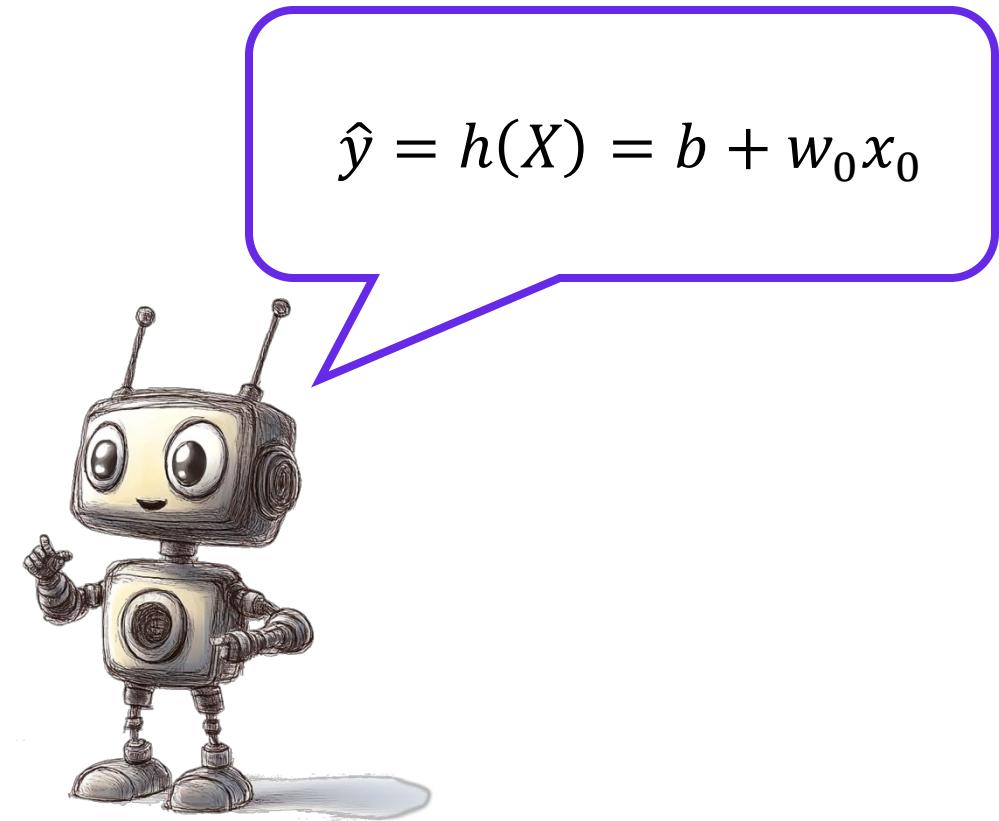
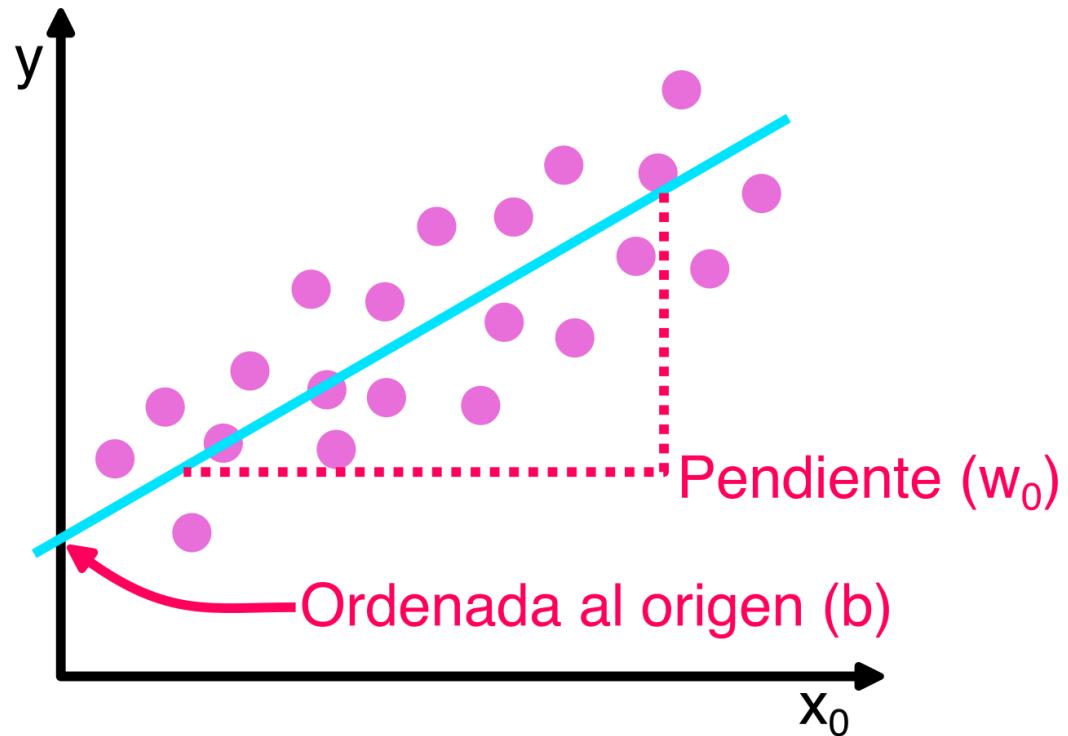
REGRESIÓN LINEAL



REGRESIÓN LINEAL

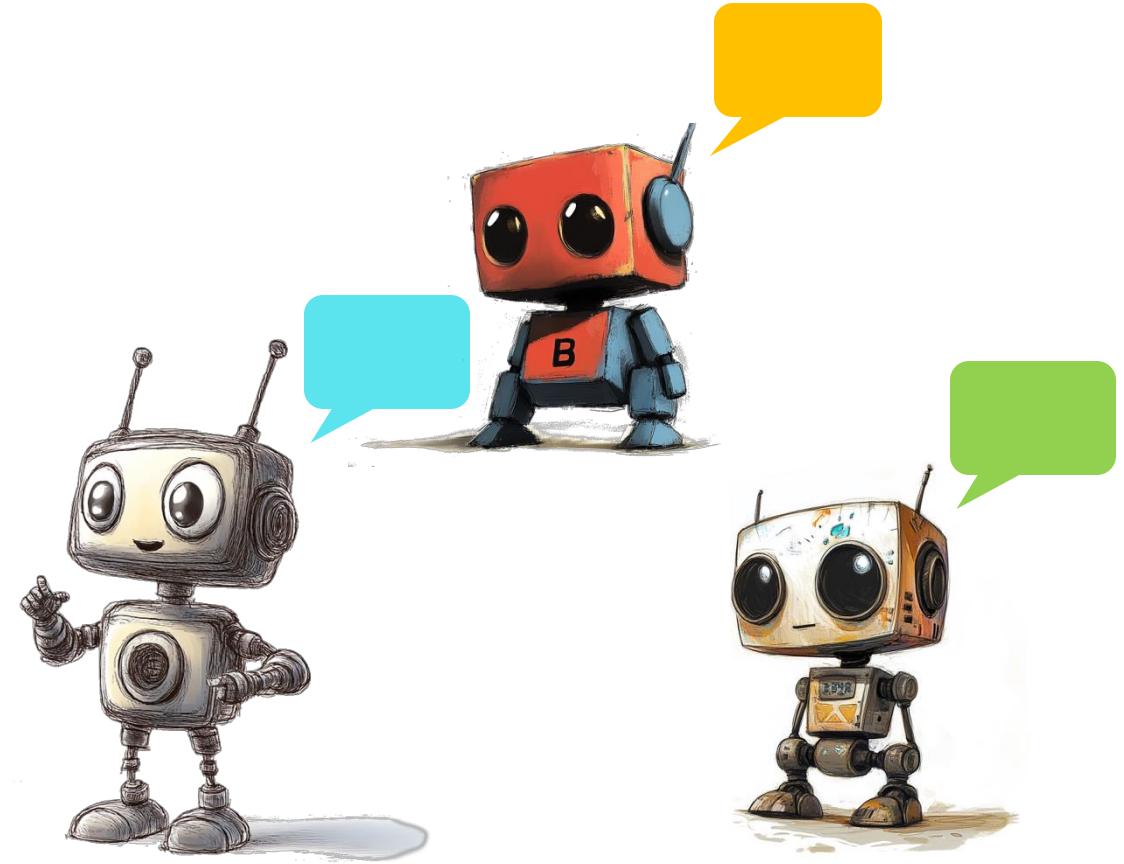
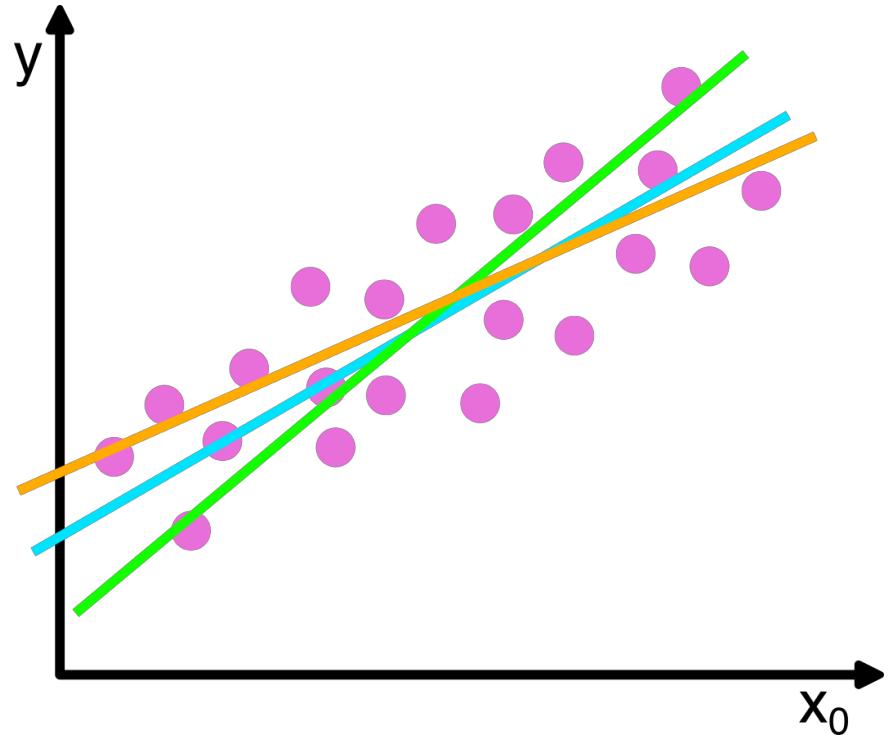


REGRESIÓN LINEAL



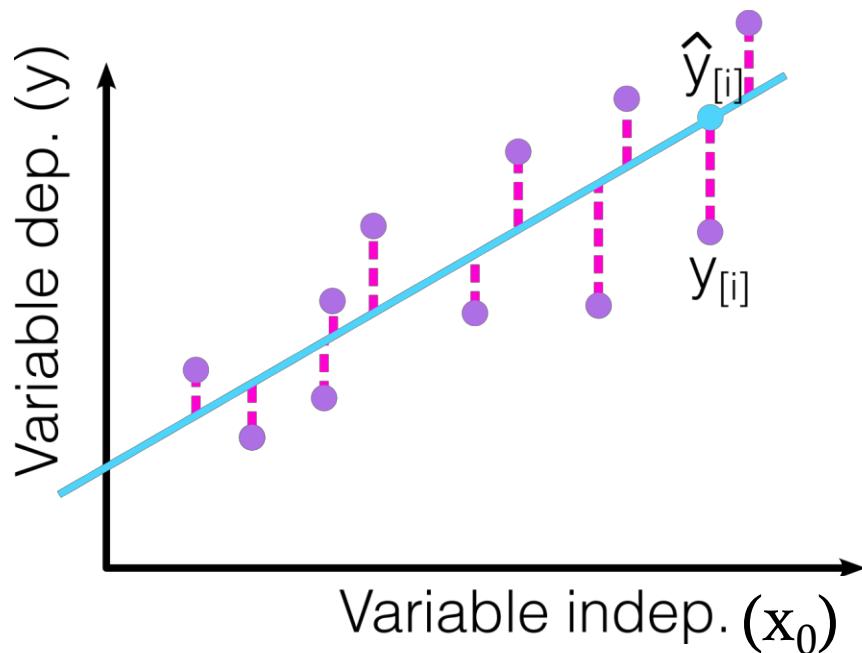
REGRESIÓN LINEAL

¿Ahora cuál recta?



REGRESIÓN LINEAL

Para encontrarla, medimos la distancia entre la recta y cada punto, a lo que llamamos **residuos**.



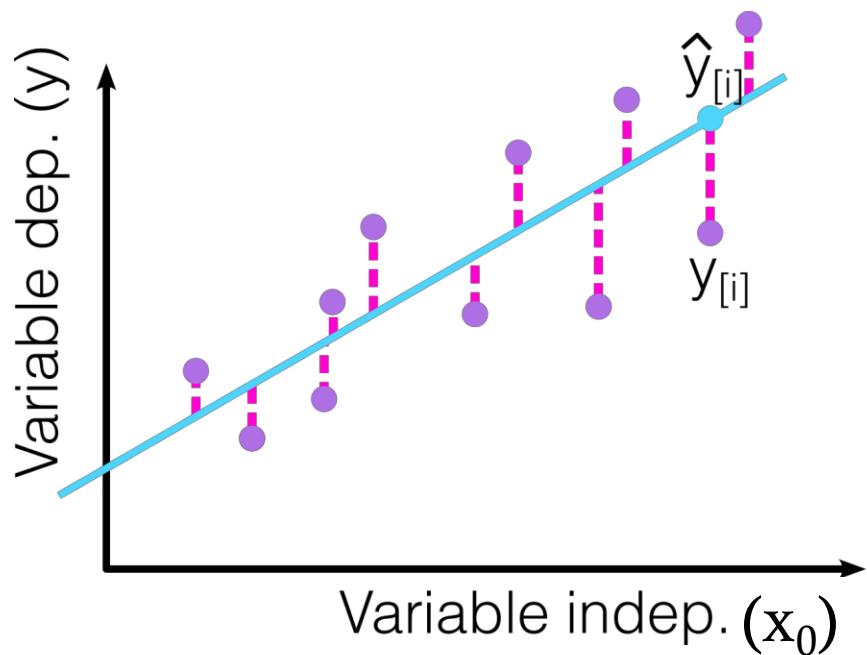
$$e_{[i]} = y_{[i]} - \hat{y}_{[i]}$$

$$y_{[i]} = \hat{y}_{[i]} + e_{[i]}$$

$$y_{[i]} = b + w_0 x_{0[i]} + e_{[i]}$$

REGRESIÓN LINEAL

Buscamos minimizar el valor de los residuos. Para lograrlo, lo hacemos minimizando la suma de los cuadrados de los **residuos**.

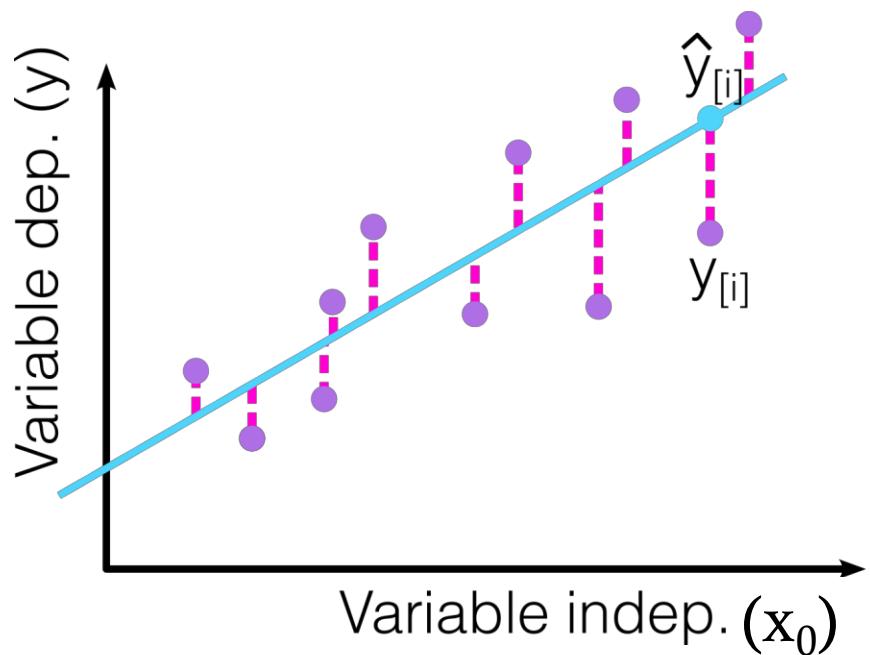


$$S_{res} = \sum_{i=0}^{N-1} (e_{[i]})^2 = \sum_{i=0}^{N-1} (y_{[i]} - b - w_0 x_{0[i]})^2$$

$$\min(S_{res}) = \min\left(\sum_{i=0}^{N-1} (e_{[i]})^2\right)$$

REGRESIÓN LINEAL

Buscamos minimizar el valor de los residuos. Para lograrlo, lo hacemos minimizando la suma de los cuadrados de los **residuos**.

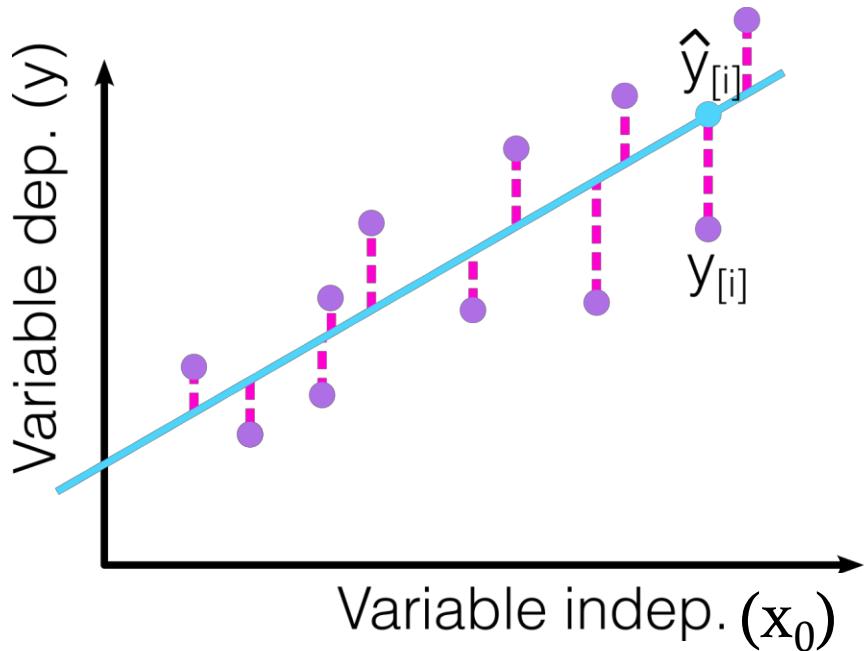


$$S_{res} = \sum_{i=0}^{N-1} (e_{[i]})^2 = \sum_{i=0}^{N-1} (y_{[i]} - b - w_0 x_{0[i]})^2$$

$$\min(S_{res}) = \min\left(\sum_{i=0}^{N-1} (e_{[i]})^2\right)$$

REGRESIÓN LINEAL

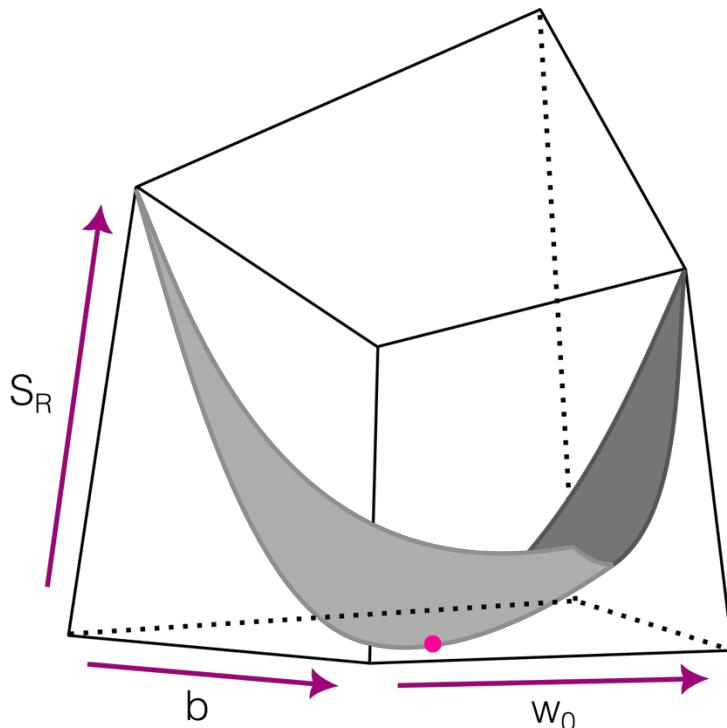
Buscamos minimizar el valor de los residuos. Para lograrlo, lo hacemos minimizando la suma de los cuadrados de los **residuos**.



Para minimizar, solo podemos ajustar los coeficientes. Lo que hacemos es seguir el **gradiente**.

$$\frac{\partial S_{res}}{\partial b} = 0 \quad \frac{\partial S_{res}}{\partial w_0} = 0$$

REGRESIÓN LINEAL



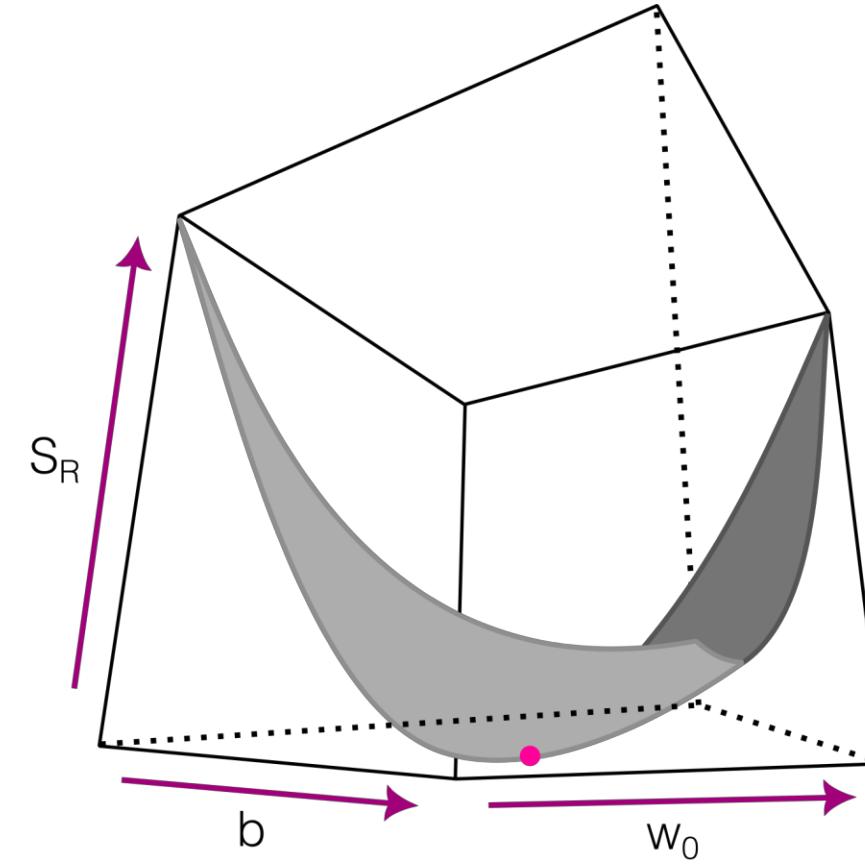
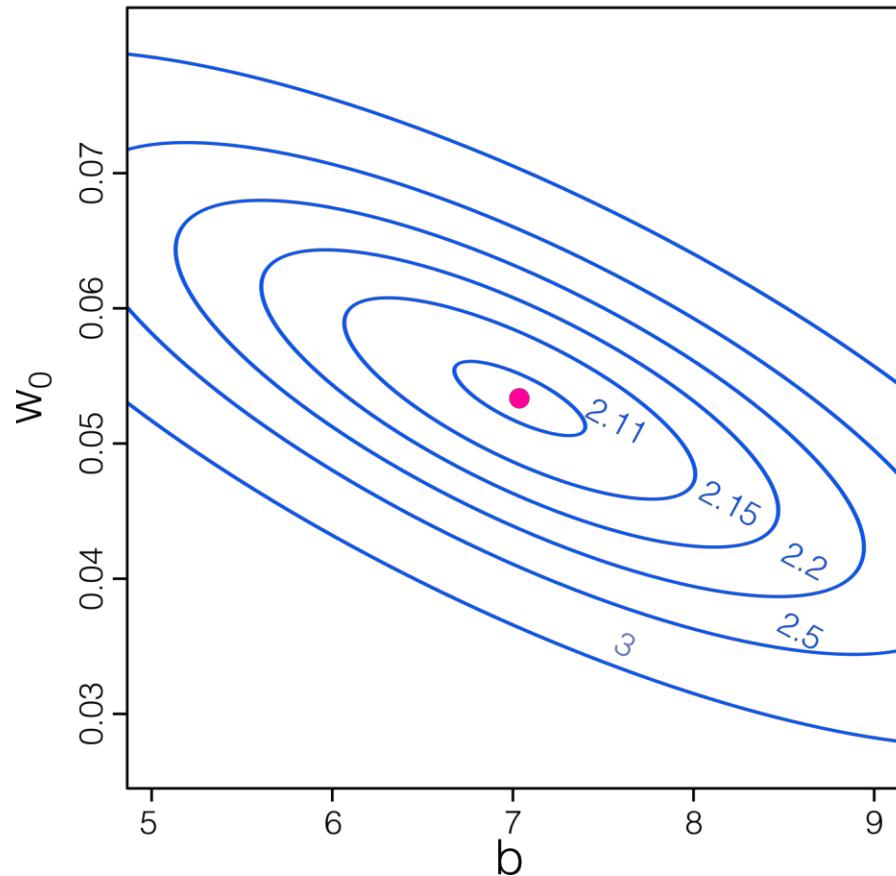
En la regresión lineal, la función es siempre convexa, es decir, siempre tiene un solo mínimo. En su forma tradicional:

$$\frac{\partial S_{res}}{\partial b} = 0 \quad \frac{\partial S_{res}}{\partial w_0} = 0$$

Si expresamos las derivadas, obtenemos un sistema de ecuaciones, llamado **ecuaciones normales**.

El problema es que, cuando tenemos muchos datos, resolver el sistema es muy difícil. ¡En esos casos, podemos usar **gradiente descendiente**!

REGRESIÓN LINEAL



REGRESIÓN LINEAL

Ajuste

¿Cómo medimos qué tan bien se ajusta una regresión a nuestros datos?

Si medimos la varianza de la variable dependiente en los datos:

$$S_T = \sum_{i=0}^{N-1} (y_{[i]} - \bar{y})^2$$

Esta varianza se puede separar en dos partes: una parte que es **atribuida al modelo** y **otra que no**:

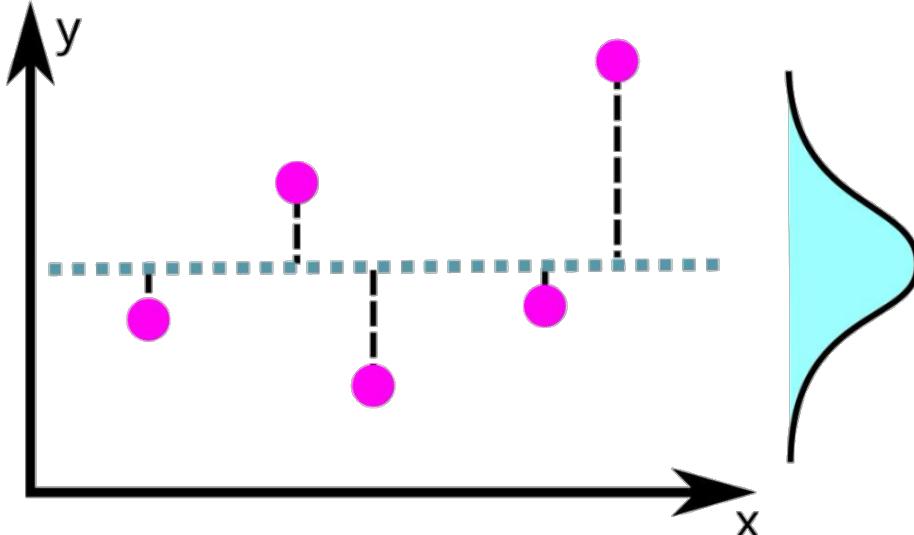
$$S_T = S_{model} + S_{res}$$
$$S_T = \sum_{i=0}^{N-1} (\hat{y}_{[i]} - \bar{y})^2 + \sum_{i=0}^{N-1} (e_{[i]})^2$$

Parte que explica el
modelo

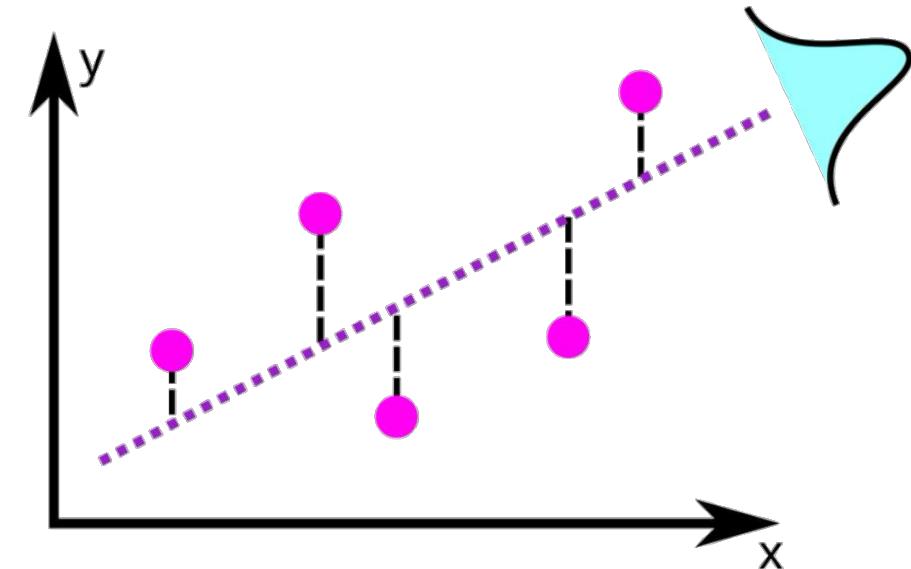
Parte que no
(residuos)

REGRESIÓN LINEAL

Ajuste



$$S_T = \sum_{i=0}^{N-1} (y_{[i]} - \bar{y})^2$$



$$S_{res} = \sum_{i=0}^{N-1} (e_{[i]})^2$$

REGRESIÓN LINEAL

Ajuste

Como métricas, podemos usar:

- El cálculo del desvío estándar residual:

$$s_{res} = \sqrt{\frac{s_{res}}{N - d - 1}} = \sqrt{\frac{1}{N - d - 1} \sum_{i=0}^{N-1} (e_{[i]})^2}$$

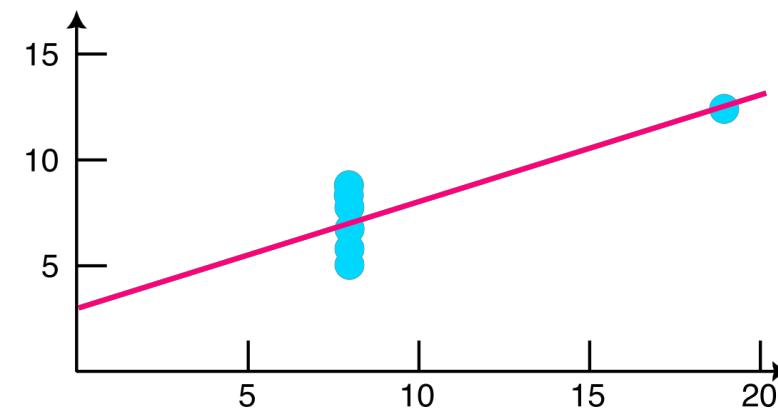
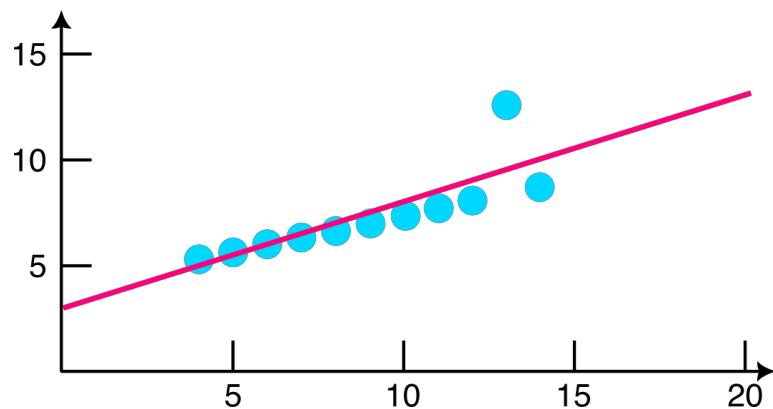
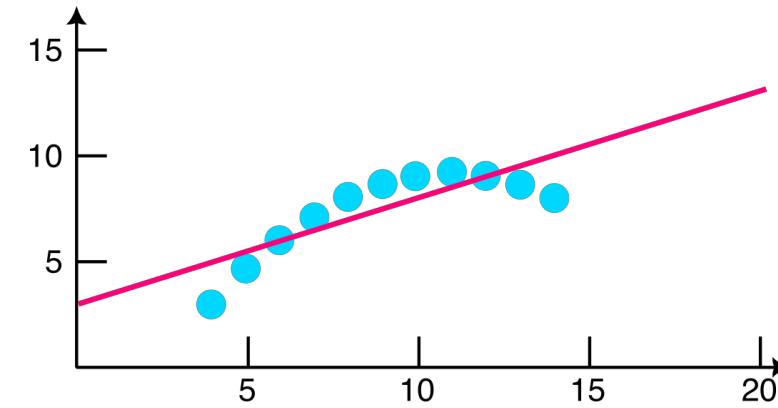
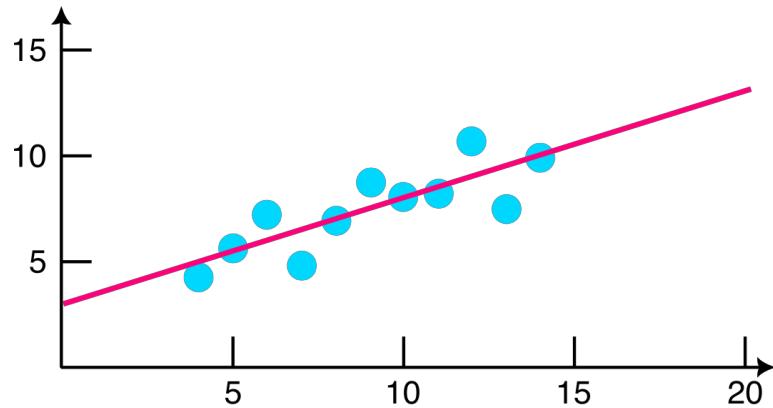
Donde d es la cantidad de *features*.

- El coeficiente de Pearson:

$$R^2 = \frac{s_{model}}{s_T} = 1 - \frac{s_{res}}{s_T}$$

REGRESIÓN LINEAL

Ajuste



REGRESIÓN LINEAL

Suposiciones

Las suposiciones que usamos para aplicar la regresión lineal son:

- **Relación lineal:** Al aplicar el modelo, muchas veces buscamos validar esta suposición.
- **Features independientes:** Los *features* de entrada de la regresión deben ser independientes entre sí.
- **Residuos:** Deben provenir de una distribución normal $N(0, \sigma^2)$ y ser independientes entre sí.

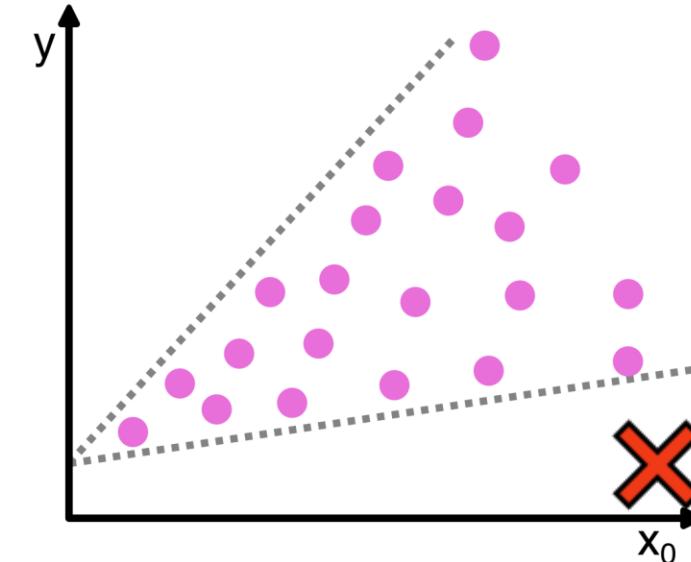
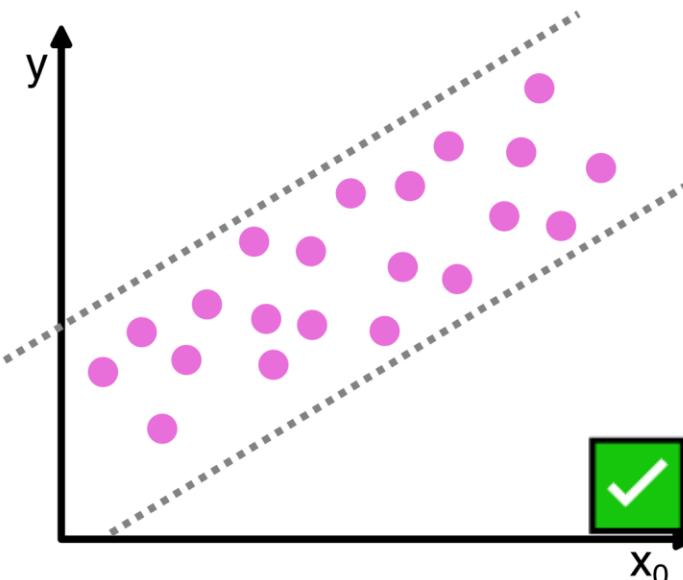
REGRESIÓN LINEAL

Suposiciones

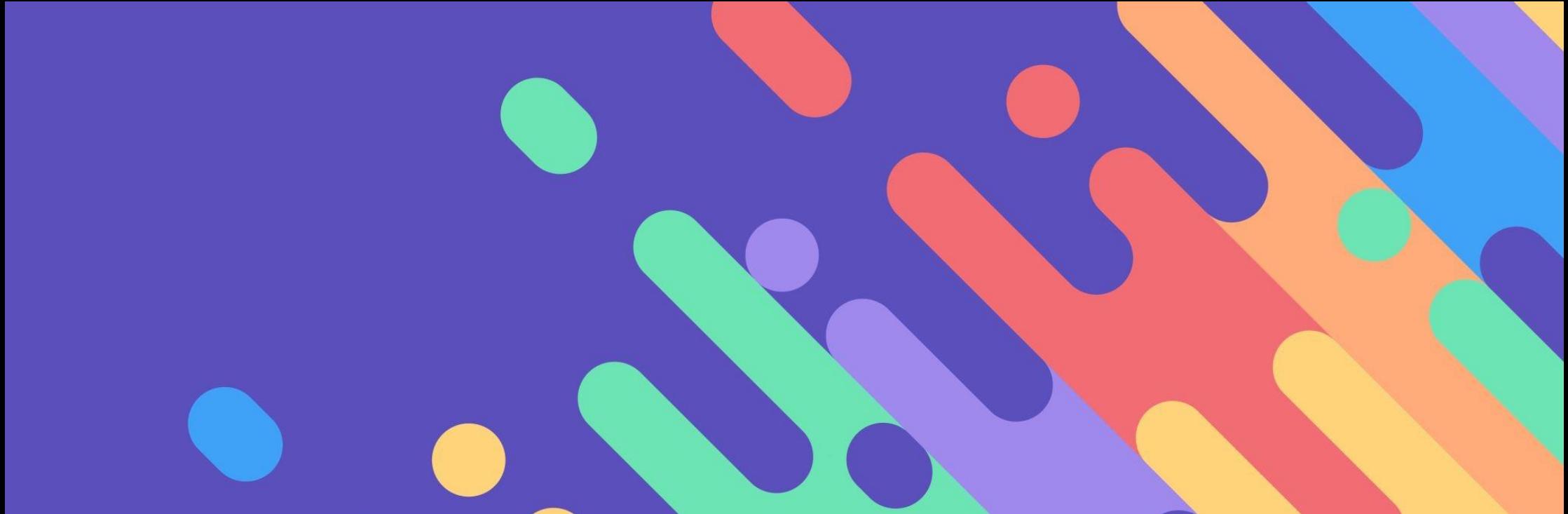
Las suposi

- **Relación** entre las variables dependientes y las independientes
- **Features** no tienen que ser independientes entre sí
- **Residuos** deben ser independientes entre sí

Homocedasticidad



r esta
n ser



TRATAMIENTO DE VARIABLES

TRATAMIENTO DE VARIABLES

Normalización o estandarización

En la regresión lineal, tenemos la multiplicación de los coeficientes por nuestras entradas:

$$\hat{y} = b + w_0x_0 + w_1x_1$$

Los coeficientes nos dan un **valor de importancia de las entradas**, pero esto es válido solo si todas las entradas están en la misma escala.

Si la variable x_0 está en rango de [1000, 3000] y x_1 en [-1, 1], los valores de w_0 y w_1 estarán en escalas diferentes, y por lo tanto, no serán comparables.

TRATAMIENTO DE VARIABLES

Normalización o estandarización

Una forma de **normalizar** es hacer que los valores estén aproximadamente entre 0 y 1, utilizando el valor máximo y el valor mínimo:

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

TRATAMIENTO DE VARIABLES

Normalización o estandarización

Una forma de **normalizar** es hacer que los valores estén aproximadamente entre 0 y 1, utilizando el valor máximo y el valor mínimo:

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Si se necesita normalizar entre -1 y 1:

$$\tilde{x} = 2 \frac{x - \min(x)}{\max(x) - \min(x)} - 1$$

TRATAMIENTO DE VARIABLES

Normalización o estandarización

Una forma de **normalizar** es hacer que los valores estén aproximadamente entre 0 y 1, utilizando el valor máximo y el valor mínimo:

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$\min(x)$
 $\max(x)$

} Set de entrenamiento

Si se necesita normalizar entre -1 y 1:

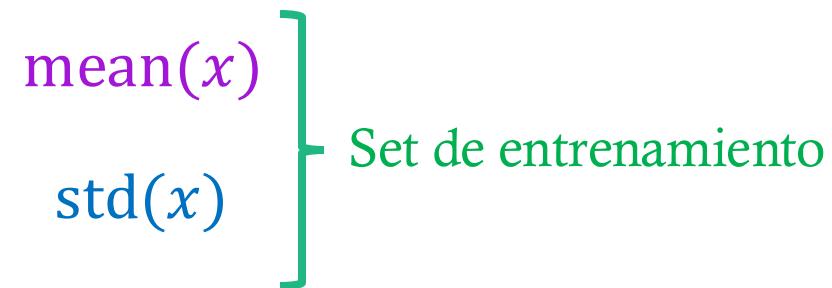
$$\tilde{x} = 2 \frac{x - \min(x)}{\max(x) - \min(x)} - 1$$

TRATAMIENTO DE VARIABLES

Normalización o estandarización

Otra forma es aplicando la **estandarización**, que es menos sensible a los **outliers**:

$$\tilde{x} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$



mean(x)
std(x)

Set de entrenamiento

A green curly brace groups the terms $\text{mean}(x)$ and $\text{std}(x)$. To the right of the brace, the text "Set de entrenamiento" is written in green.

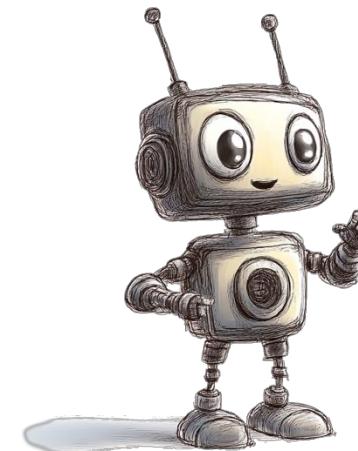
TRATAMIENTO DE VARIABLES

Variables Dummies

La regresión lineal utiliza variables numéricas para predecir un valor. ¿Cómo podemos usar **variables categóricas**?

Para poder usarlas, debemos transformarlas en *numéricas mediante alguna codificación*.

En el caso de **variables categóricas ordinales**, podemos asociarles un número:



- | |
|---------------|
| ★ ★ ★ ★ ★ = 1 |
| ★ ★ ★ ★ ☆ = 2 |
| ★ ★ ★ ☆ ☆ = 3 |
| ★ ★ ★ ☆ ☆ = 4 |
| ★ ★ ★ ★ ☆ = 5 |

TRATAMIENTO DE VARIABLES

Variables Dummies

Si tenemos **casos nominales** no podemos asociarles números, ya que, al hacerlo, estableceríamos un orden.

Para este tipo de variable existen diferentes codificaciones, y una de ellas es el **one-hot encoding**.



Ale = [1, 0, 0]

Honey = [0, 1, 0]

Stout = [0, 0, 1]

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	País
80	180	Argentina
83	177	Chile
75	169	Chile
68	155	Argentina
95	199	Perú

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	País	Argentina	Chile	Perú
80	180	Argentina	1	0	0
83	177	Chile	0	1	0
75	169	Chile	0	1	0
68	155	Argentina	1	0	0
95	199	Perú	0	0	1

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	Argentina	Chile	Perú
80	180	1	0	0
83	177	0	1	0
75	169	0	1	0
68	155	1	0	0
95	199	0	0	1

TRATAMIENTO DE VARIABLES

Variables Dummies

El **one-hot encoding** nos genera un nuevo atributo por categoría, pero esto puede generar *una trampa*

Si vemos el ejemplo, las tres variables que estamos usando están 100% correlacionadas entre sí:

$$\hat{y} = b + w_0 x_{peso} + w_1 x_{altura} + w_2 x_{arg} + w_3 x_{chile} + w_4 x_{peru}$$

$$x_{peru} = 1 - x_{arg} - x_{chile}$$

$$\hat{y} = b + w_0 x_{peso} + w_1 x_{altura} + w_2 x_{arg} + w_3 x_{chile} + w_4(1 - x_{arg} - x_{chile})$$

$$\hat{y} = (b + w_4) + w_0 x_{peso} + w_1 x_{altura} + (w_2 - w_4)x_{arg} + (w_3 - w_4)x_{chile}$$

Para solucionar esto, debemos eliminar siempre una columna para romper la trampa.

TRATAMIENTO DE VARIABLES

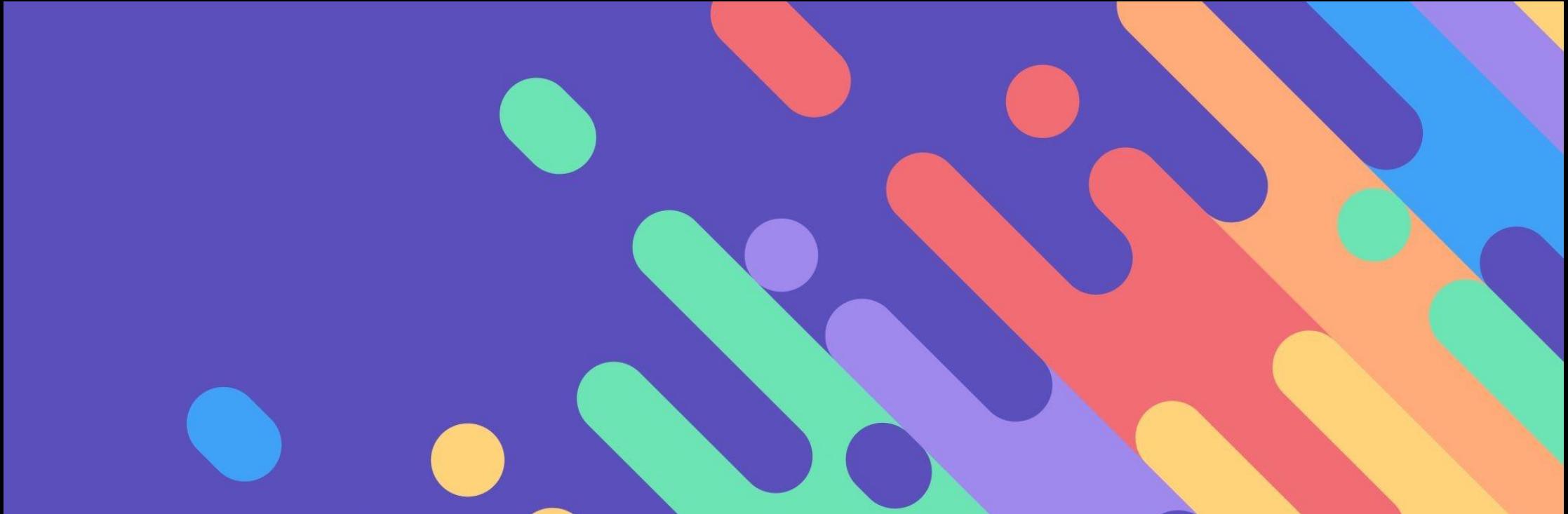
Variables Dummies

Peso	Altura	Argentina	Chile	Perú
80	180	1	0	0
83	177	0	1	0
75	169	0	1	0
68	155	1	0	0
95	199	0	0	1

TRATAMIENTO DE VARIABLES

Variables Dummies

Peso	Altura	Argentina	Chile
80	180	1	0
83	177	0	1
75	169	0	1
68	155	1	0
95	199	0	0



MÉTRICAS DE EVALUACIÓN

MÉTRICAS DE EVALUACIÓN

El conjunto de datos de evaluación se utiliza para evaluar qué tan bien se entrenó el algoritmo con el conjunto de datos de entrenamiento.

¿Pero cómo evaluamos?

- **El coeficiente de Pearson (R^2)**

Podemos usar métricas más generales, que midan el error en variables numéricas.

MÉTRICAS DE EVALUACIÓN

Error absoluto medio (MAE)

El **error absoluto medio (MAE)** es el cálculo del valor absoluto del residuo para cada punto de datos.

Luego, tomamos el promedio de todos estos residuos.

$$MAE = \frac{1}{N} \sum_{i=0}^{N-1} |y_{[i]} - \hat{y}_{[i]}|$$

MÉTRICAS DE EVALUACIÓN

Error cuadrático medio (MSE)

El **error cuadrático medio (MSE)** es similar al **MAE**, pero ahora calculamos el cuadrado de los residuos.

$$MSE = \frac{1}{N} \sum_{i=0}^{N-1} (y_{[i]} - \hat{y}_{[i]})^2$$

Un detalle importante son los residuos grandes (**outliers**). En esta métrica, su impacto es mayor que en el **MAE**.

MÉTRICAS DE EVALUACIÓN

Raíz cuadrada del error cuadrático medio (RMSE)

Si al **MSE** le calculamos la raíz cuadrada, obtenemos una métrica llamada **RMSE**, que tiene la misma unidad que la salida original, a diferencia del **MSE**, que no la tiene.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (y_{[i]} - \hat{y}_{[i]})^2}$$

MÉTRICAS DE EVALUACIÓN

Outliers

Los valores atípicos son un tema de constante discusión. *¿Se deben incluir o no?*

La respuesta dependerá del problema en particular, de los datos disponibles y de las consecuencias que tenga el considerar o no esos valores.

Si quiero tenerlos en cuenta a la hora de comparar modelos, me convendrá usar **MSE**. En cambio, si quiero reducir su importancia, puedo usar **MAE**.

MÉTRICAS DE EVALUACIÓN

Error absoluto porcentual medio (MAPE)

El **error absoluto porcentual medio (MAPE)** es el cálculo del error **MAE**, pero escalado al valor verdadero, por lo que el resultado es porcentual:

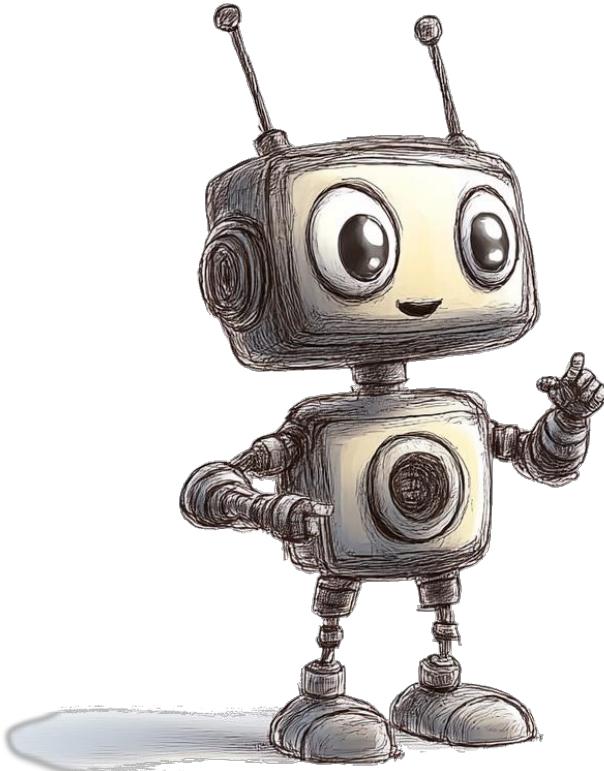
$$MAPE = \frac{100\%}{N} \sum_{i=0}^{N-1} \left| \frac{y_{[i]} - \hat{y}_{[i]}}{y_{[i]}} \right|$$

No es una métrica ideal porque es susceptible a errores numéricos. No puede calcularse cuando $y_{[i]}$ es igual a cero. Además, tiene sesgo cuando la predicción subestima:

$$\begin{aligned} n &= 1 & \hat{y} &= 10 & y &= 20 \\ MAPE &= 50\% \end{aligned}$$

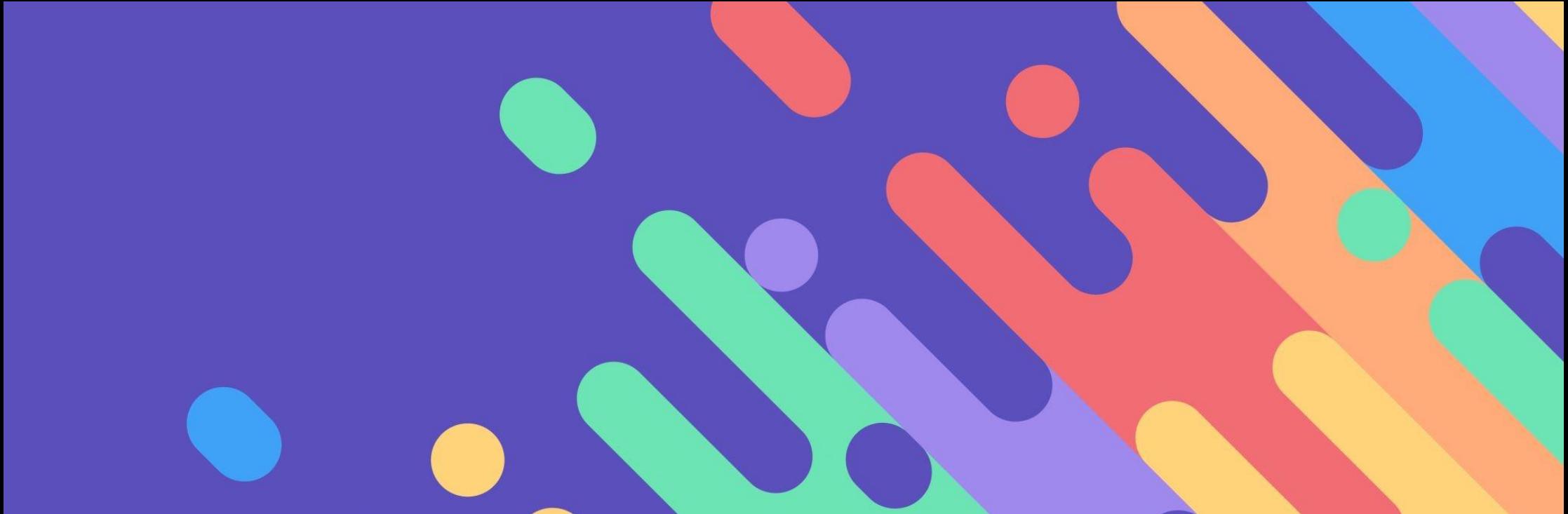
$$\begin{aligned} n &= 1 & \hat{y} &= 20 & y &= 10 \\ MAPE &= 100\% \end{aligned}$$

REGRESIÓN



Los temas que vamos a ver en este video son:

- Construcción de un modelo
 - Eliminación hacia atrás
 - Selección hacia adelante
 - Eliminación bidireccional
- Regresión de Ridge
- Regresión de Lasso



CONSTRUCCIÓN DE UN MODELO

CONSTRUCCIÓN DE UN MODELO

Cuando construimos un modelo de regresión múltiple, ¿cómo elegimos los atributos que formarán parte del modelo?

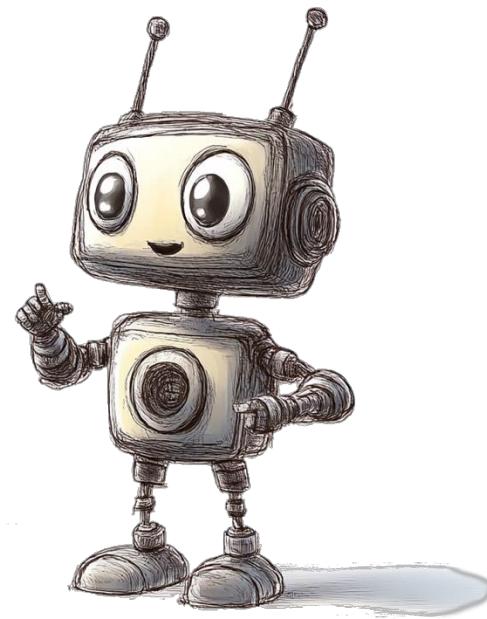
Ver la correlación entre variables es un primer paso, pero surge la pregunta: si dos variables están correlacionadas, ¿cuál de las dos descartamos?

Por lo tanto, existen diferentes métodos para construir un modelo.

CONSTRUCCIÓN DE UN MODELO

Podemos mencionar 4 formas:

- Exhaustivo
- Eliminación hacia atrás
- Selección hacia adelante
- Eliminación bidireccional



CONSTRUCCIÓN DE UN MODELO

Eliminación hacia atrás

- Se comienza con un modelo completo que incluye todas las variables.
- Luego, se eliminan de **forma greedy** las variables de entrada que menos "*aportan*" al modelo, una por vez.
- El proceso continúa hasta que eliminar más variables no mejora significativamente el modelo.
- Este proceso se puede realizar utilizando alguna métrica que nos mida la información que aporta cada variable.

CONSTRUCCIÓN DE UN MODELO

Selección hacia adelante

- Comienza con un modelo *vacio* que solo incluye la ordenada al origen:
$$\bar{y} = \frac{1}{N} \sum_{i=0}^{N-1} y_{[i]}.$$
- Luego, se agregan las variables que más aportan al modelo (usando el criterio de ajuste), una por vez.
- El proceso termina cuando agregar más variables no mejora significativamente el modelo.

CONSTRUCCIÓN DE UN MODELO

Eliminación bidireccional

- Es, en esencia, la selección hacia adelante, pero con la posibilidad de eliminar variables en cada iteración si se observa correlación entre ellas.

CONSTRUCCIÓN DE UN MODELO

¿Cómo sabemos el aporte de cada atributo?

- **Bondad de ajuste:** Se realiza una prueba de hipótesis para determinar si el coeficiente de una entrada en particular es cero. Luego, se evalúa el valor p. Un valor p bajo ($< 0,05$) indica que se puede rechazar la hipótesis nula, lo que sugiere que cambios en esta variable probablemente generen cambios en la respuesta.

CONSTRUCCIÓN DE UN MODELO

¿Cómo sabemos el aporte de cada atributo?

- **Coeficiente de Pearson ajustado:** Es el cual es el R^2 pero penalizando la complejidad del modelo:

$$R_{adj}^2 = 1 - (1 - R^2) \frac{N - 1}{N - d - 1}$$

Donde N es número de observaciones y d es el número de atributos.

CONSTRUCCIÓN DE UN MODELO

¿Cómo sabemos el aporte de cada atributo?

- **Criterio de Información de Aikake (AIC):** El AIC maneja un equilibrio entre la bondad de ajuste del modelo y la complejidad de este. En otras palabras, el AIC aborda tanto el riesgo de sobreajuste como el riesgo de subajuste.

$$AIC = 2d - 2 \ln(\hat{L})$$

El valor de AIC debe ser más bajo para indicar un mejor modelo.

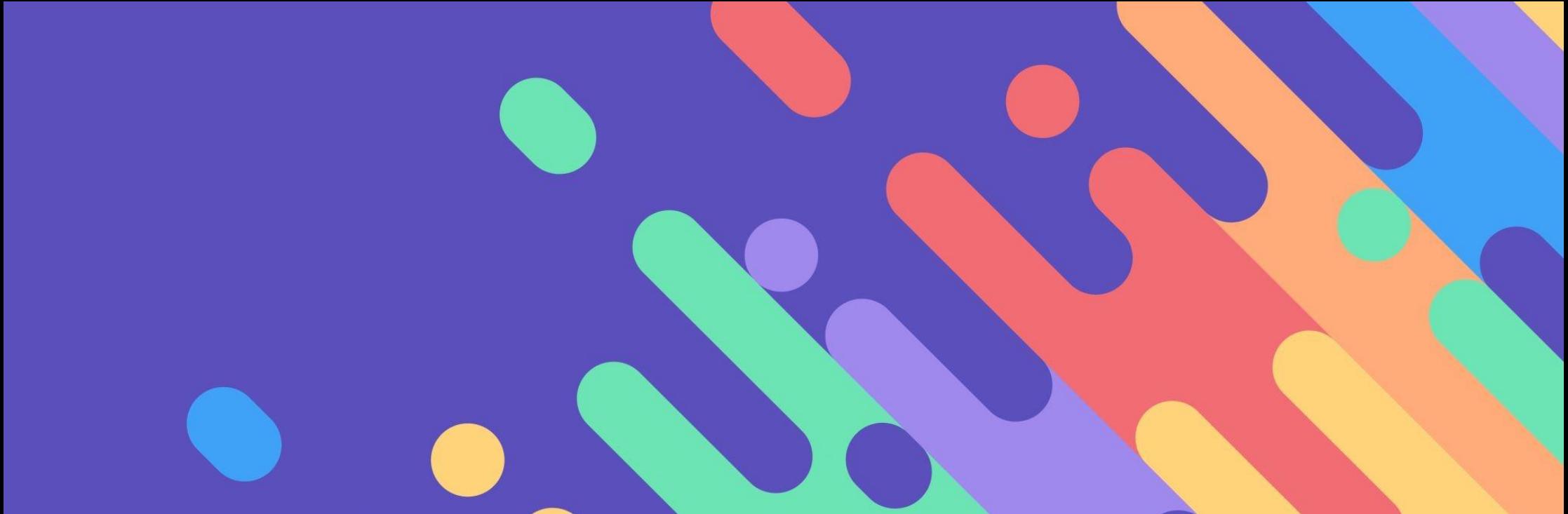
CONSTRUCCIÓN DE UN MODELO

¿Cómo sabemos el aporte de cada atributo?

- **Criterio de información bayesiano (BIC):** Se basa en el principio de la navaja de Occam, que establece que es preferible el modelo más simple que explique los datos. A diferencia del AIC, el BIC penaliza más al modelo por su complejidad.

$$BIC = d \ln(N) - 2\ln(\hat{L})$$

El valor de BIC debe ser más bajo para indicar un mejor modelo.



REGRESIÓN LASSO Y RIDGE

REGRESIÓN DE RIDGE Y LASSO

¿Cómo sabemos el aporte de cada atributo?

Con los métodos de regularización, podemos ajustar un modelo que contenga todos los atributos utilizando una técnica que restrinja o regularice las estimaciones de los coeficientes.

Puede que no sea inmediatamente obvio por qué tal restricción debería mejorar el ajuste, pero resulta que **reducir las estimaciones de los coeficientes** puede **disminuir significativamente su varianza**.

Las dos técnicas más conocidas para reducir los coeficientes de regresión a cero son la regresión de **Ridge** y la regresión de **Lasso**.

REGRESIÓN DE RIDGE

En la regresión lineal, vimos que se buscaban los coeficientes que minimizaban la suma de los residuos al cuadrado. La **regresión de Ridge** es muy similar, pero con la diferencia de que los coeficientes se estiman minimizando una cantidad diferente:

$$\sum_{i=0}^{N-1} (y_{[i]} - b - \mathbf{W}^T \mathbf{X}_{[i]})^2 + \alpha \sum_{j=0}^{d-1} w_j^2$$

Donde α es un hiperparámetro de ajuste.

REGRESIÓN DE RIDGE

En esta regresión, se buscan los coeficientes que minimizan S_{res} . Sin embargo, el segundo término:

$$\alpha \sum_{j=0}^{d-1} w_j^2$$

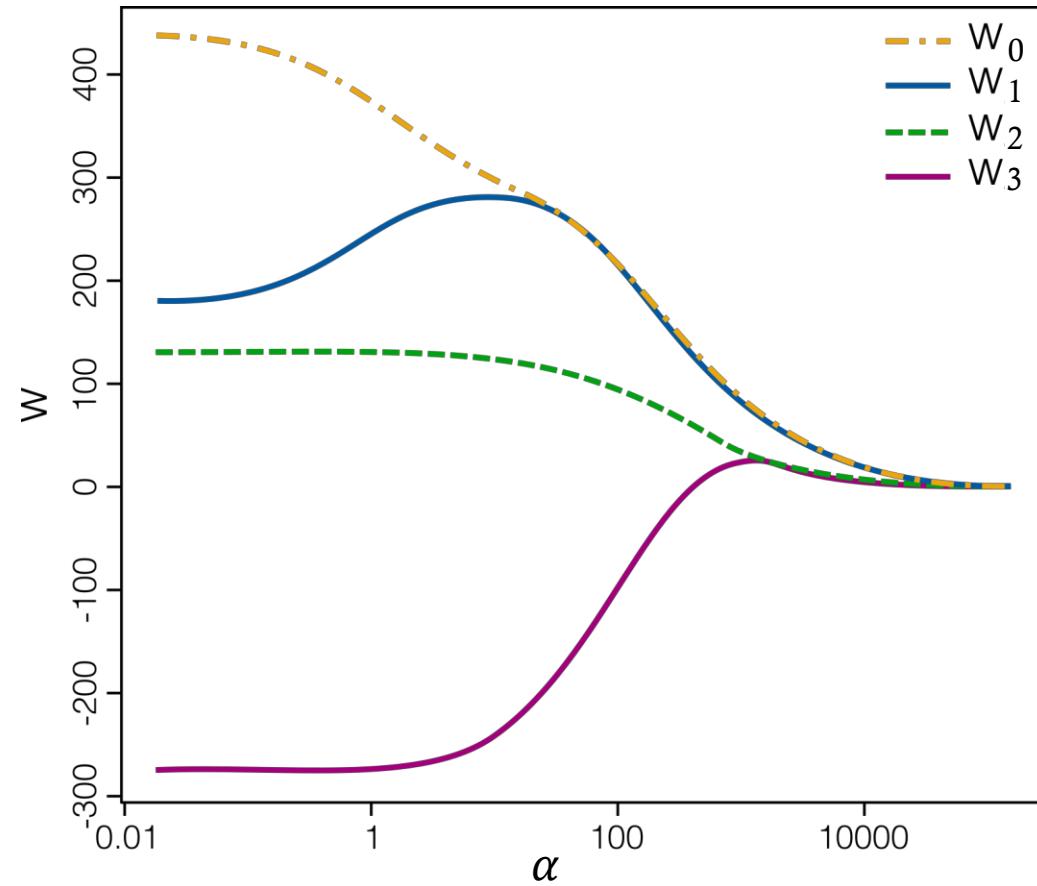
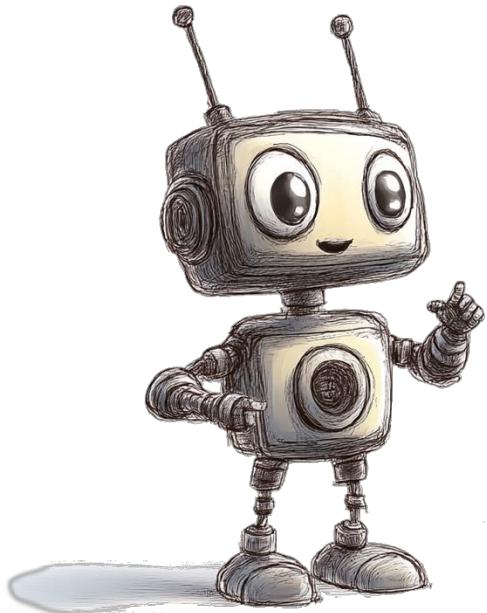
Se conoce como el término de **penalización por encogimiento**. Este término es pequeño cuando los coeficientes están cerca de cero.

REGRESIÓN DE RIDGE

Nótese que la penalización **no afecta la ordenada al origen b** . Si α tiende a ∞ , todos los coeficientes se vuelven cero, y la regresión queda de la siguiente forma:

$$\hat{y} = b = \frac{1}{N} \sum_{i=0}^{N-1} y_{[i]}$$

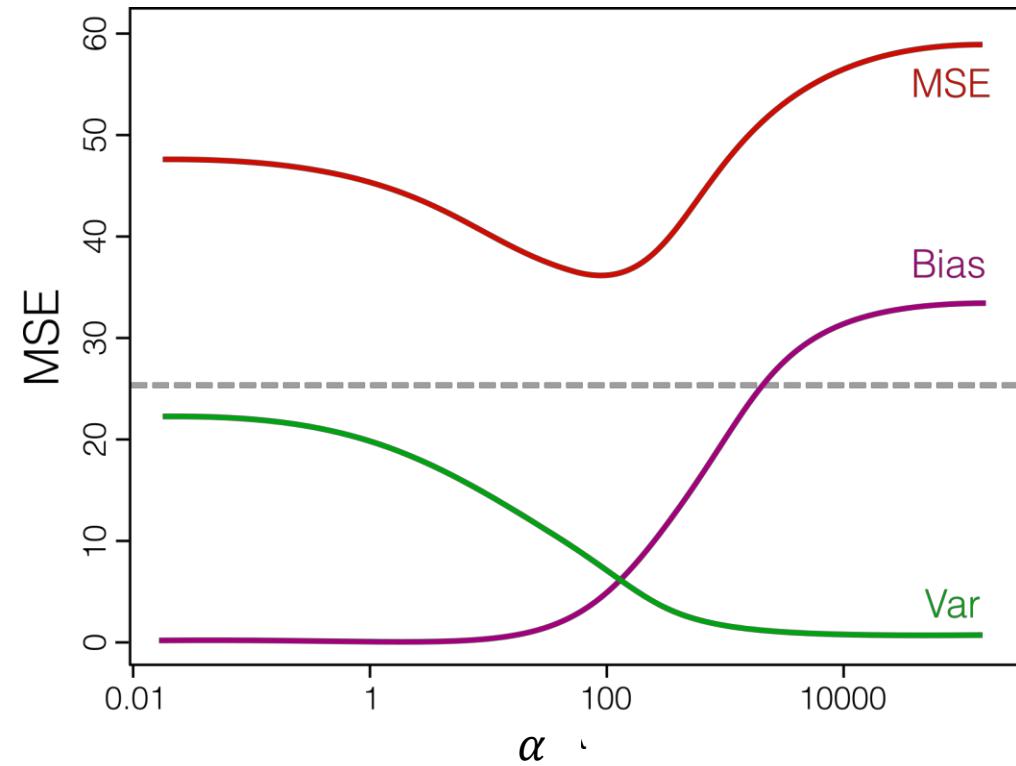
REGRESIÓN DE RIDGE



REGRESIÓN DE RIDGE

¿Para qué nos sirve?

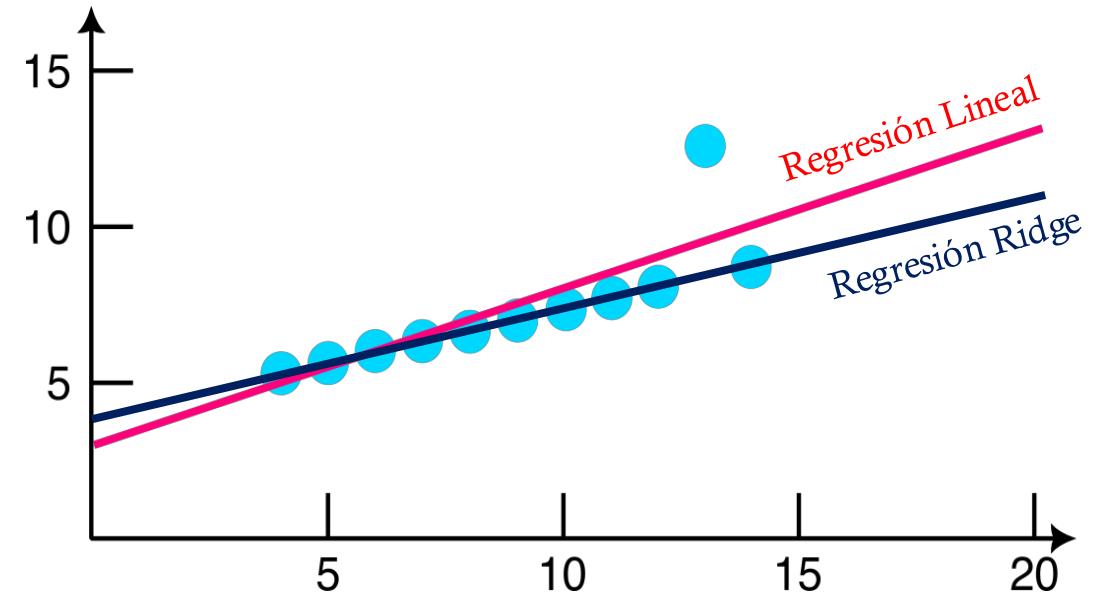
La ventaja de la regresión de Ridge sobre la regresión lineal de mínimos cuadrados radica en el equilibrio entre sesgo y varianza. A medida que α aumenta, la **flexibilidad del ajuste disminuye**, lo que lleva a una **menor varianza**, pero a un **mayor sesgo**.



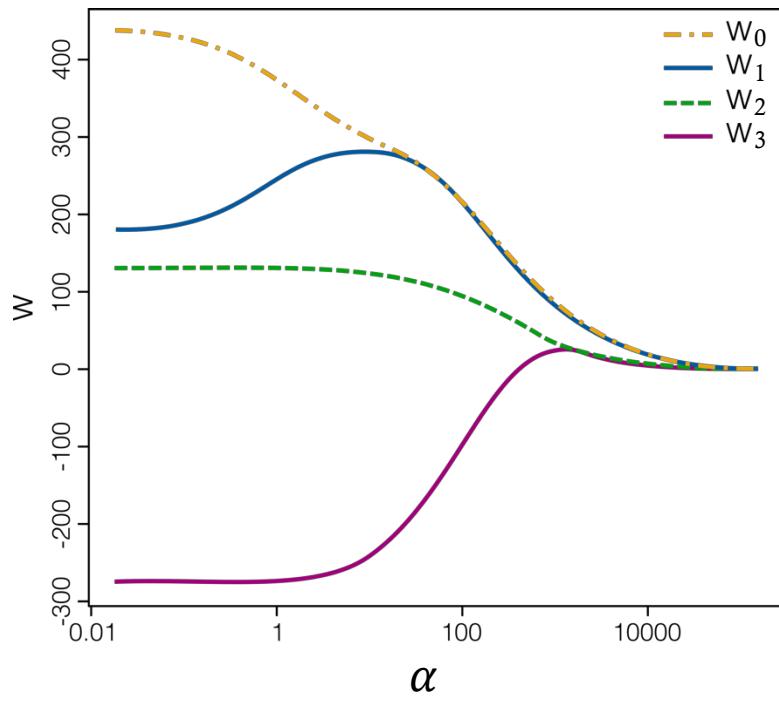
REGRESIÓN DE RIDGE

¿Para qué nos sirve?

En general, cuando la verdadera relación es lineal, la regresión lineal tiende a tener mucha varianza. Esto ocurre principalmente cuando el *número de observaciones es cercano al número de coeficientes*.



REGRESIÓN DE LASSO



La regresión de Ridge, a priori, nos parece interesante para hacer una selección de modelo, ya que, al ajustar α , podemos ver si algún coeficiente se acerca a cero.

El problema es que los coeficientes se reducen hacia cero, pero no se hacen exactamente cero, a menos que α sea infinito. Por lo tanto, no podemos eliminar atributos.

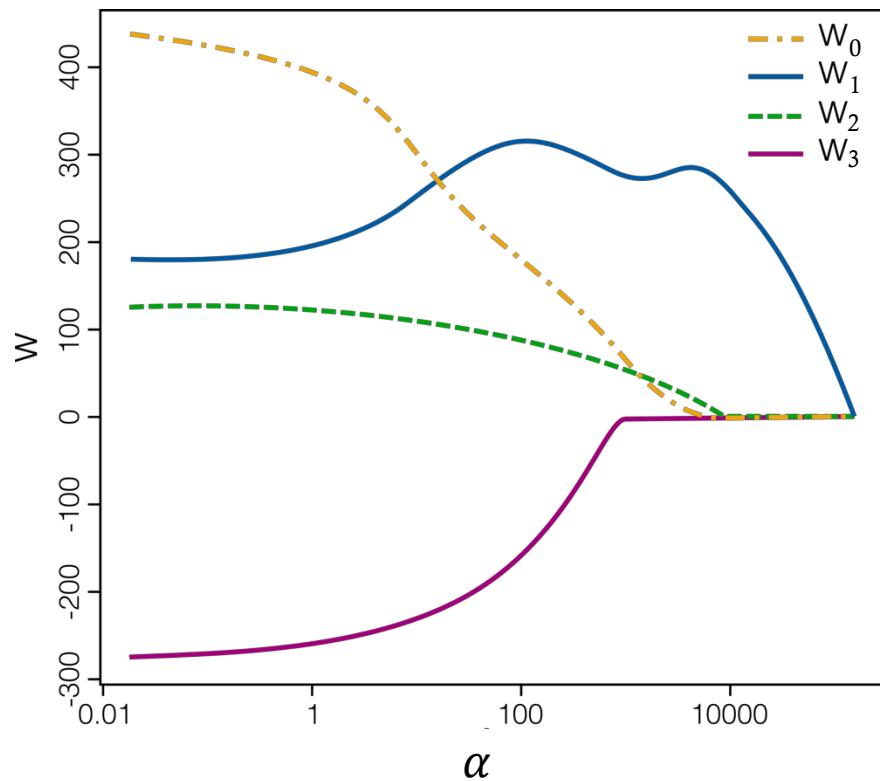
REGRESIÓN DE LASSO

La regresión de Lasso cubre esta desventaja:

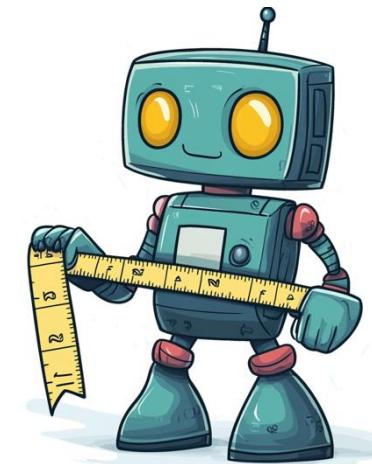
$$\sum_{i=0}^{N-1} (y_{[i]} - b - \mathbf{W}^T \mathbf{X}_{[i]})^2 + \alpha \sum_{j=0}^{d-1} |w_j|$$

La regresión de Lasso utiliza una **penalización L1**, mientras que Ridge utiliza una **penalización L2**.

REGRESIÓN DE LASSO



Esta regresión, cuando α crece, hace que algunos coeficientes se conviertan exactamente en cero. *Por lo tanto, Lasso realiza una selección de atributos.*



REGRESIÓN DE LASSO

¿Para qué nos sirve?

Para entender por qué esto ocurre, debemos reescribir las regresiones de una forma equivalente:

Regresión Lasso

$$\min \left\{ \sum_{i=0}^{N-1} (y_{[i]} - b - \mathbf{W}^T \mathbf{X}_{[i]})^2 \right\} \quad \text{sujeto a } \sum_{j=0}^{d-1} |w_j| \leq s$$

Regresión Ridge

$$\min \left\{ \sum_{i=0}^{N-1} (y_{[i]} - b - \mathbf{W}^T \mathbf{X}_{[i]})^2 \right\} \quad \text{sujeto a } \sum_{j=0}^{d-1} w_j^2 \leq s$$

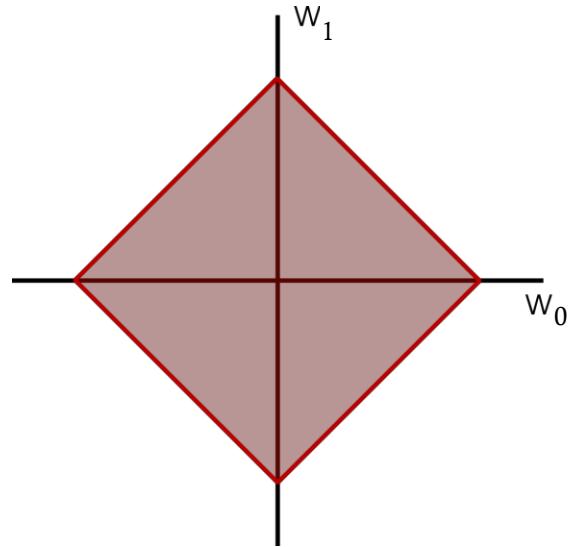
REGRESIÓN DE LASSO

¿Para qué nos sirve?

Veamos el efecto de la penalización en un caso con 2 atributos ($d= 2$):

Regresión Lasso

$$|w_0| + |w_1| \leq s$$



Regresión Ridge

$$w_0^2 + w_1^2 \leq s$$

