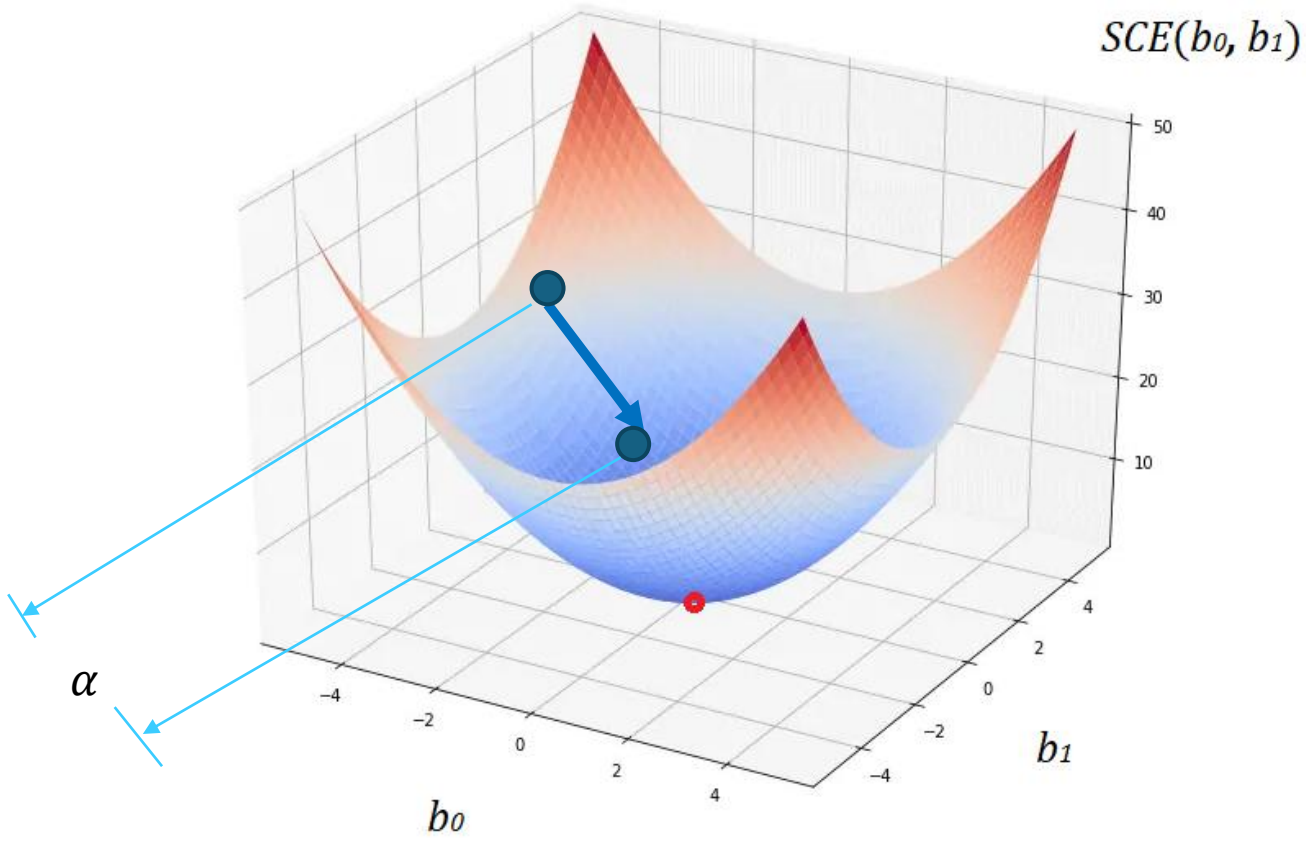
Regresión lineal con gradiente descendiente

Dra. Consuelo Varinia García Mendoza

Gradiente descendiente por lotes (BGD)

- Tamaño de paso = $\alpha \in (0,1)$
- Dirección -> derivada
- No. de pasos (iteraciones)



BGD

$$w_i = w_i - \alpha \frac{\partial f(w_i)}{\partial w_i}$$

$$w_i = w_i - 2\alpha \sum_{j=0}^{m-1} \left(w_i x_{j,i} - y_j \right) \cdot x_{j,i}$$

$$w_i = w_i - 2\alpha \cdot sum \left(\left(w_i \begin{bmatrix} x_{0,i} \\ x_{1,i} \\ \vdots \\ x_{m-1,i} \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{m-1} \end{bmatrix} \right) \cdot \begin{bmatrix} x_{0,i} \\ x_{1,i} \\ \vdots \\ x_{m-1,i} \end{bmatrix} \right)$$

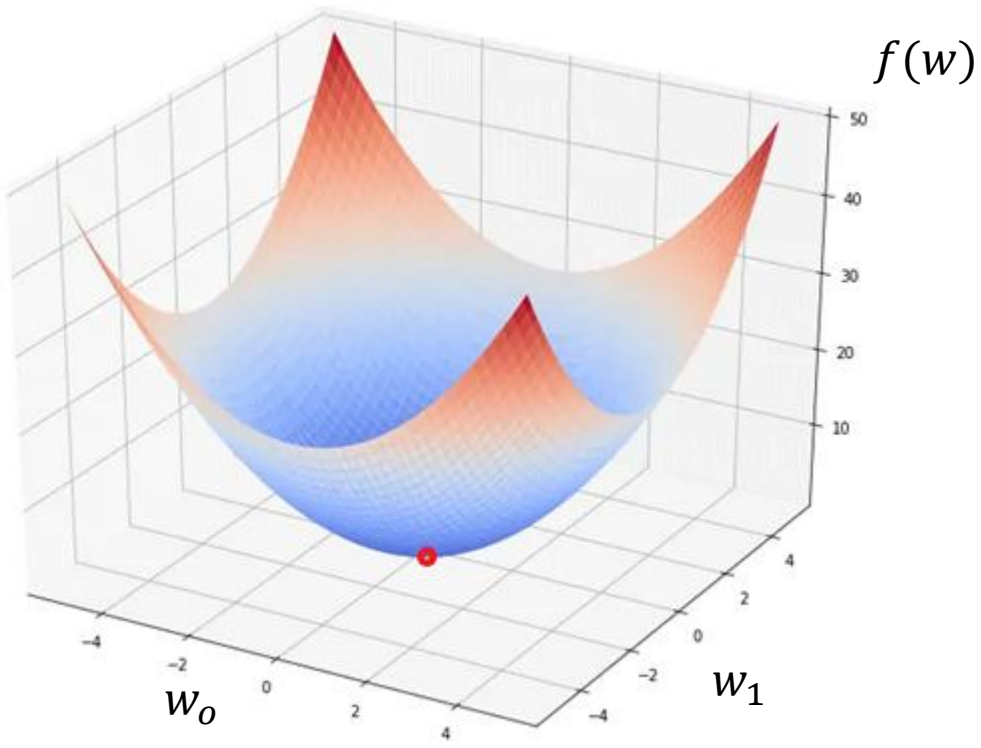
Limitaciones del algoritmo de BGD

- Permite ajustar los pesos de forma iterativa
- En cada iteración se calculan los pesos de cada característica y del error
- Considera todos los ejemplos en su cálculo
- Para conjuntos de datos grandes esto puede ser muy costoso y BGD se vuelve lento

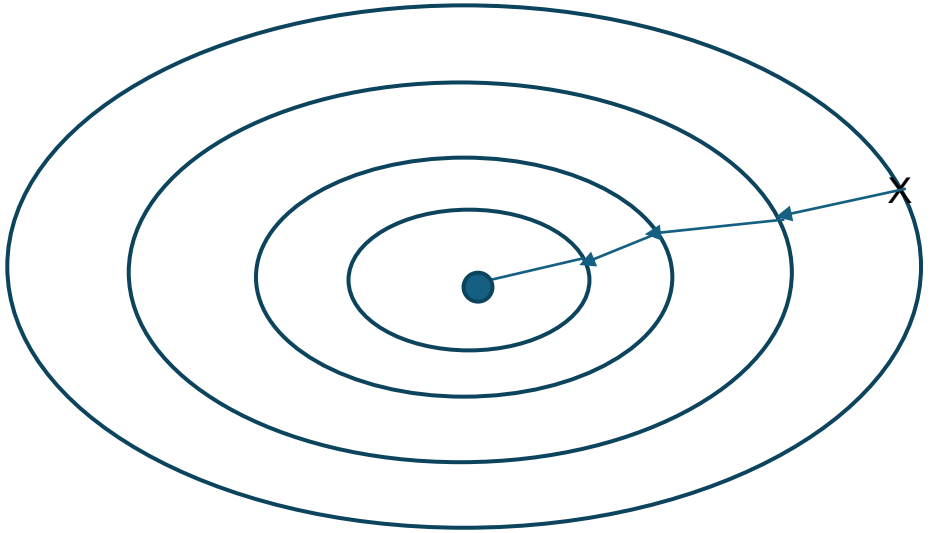
Gradiente descendente estocástico (SGD)

- Existe una variante del algoritmo BGD que considera sólo un ejemplo para el ajuste de pesos en cada iteración. Eligiendo de manera aleatoria este ejemplo es decir se establece r de manera aleatoria en cada iteración
- Con este único ejemplo se ajustan los pesos y se intenta una aproximación al óptimo
- Este algoritmo reduce el número de operaciones que se realizan en cada iteración por lo que puede manejar grandes cantidades de datos
- Sin embargo, se debe considerar que no siempre se encontrarán los pesos óptimos, pero si una buena aproximación

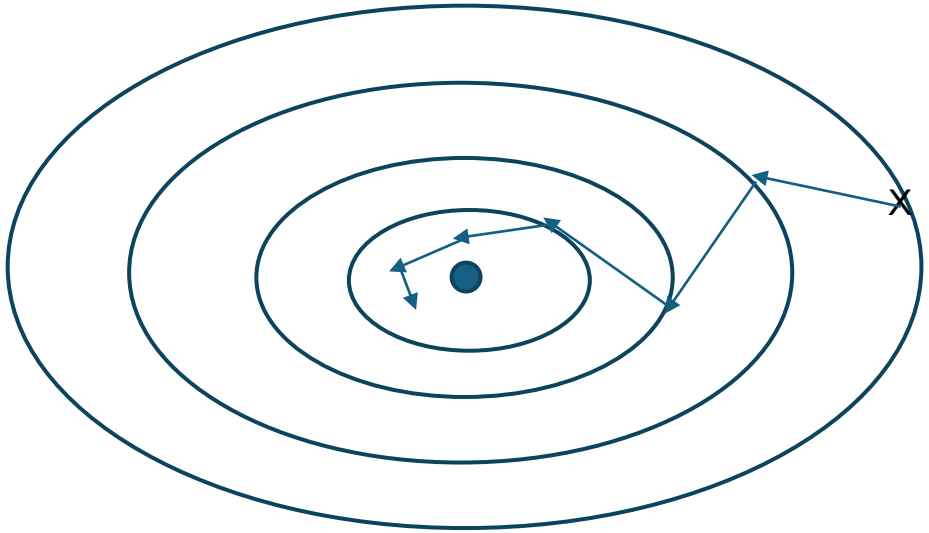
$$w_k = w_k - 2\alpha\left(w_k x_{r,k} - y_r\right) \cdot x_{r,k}$$



BGD



SGD



Implementación de BGD o SGD

- Funciones de perdida
  - Suma de errores al cuadrado

$$f(w_k) = \sum_{r=0}^{e-1} \left(w_k x_{r,k} - y_r \right)^2$$

  - Error cuadrático medio

$$f(w_k) = \frac{1}{e} \sum_{r=0}^{e-1} \left(w_k x_{r,k} - y_r \right)^2$$

Regresión lineal

- Mínimos Cuadrados Ordinarios (OLS)
  - Método analítico

$$\frac{\partial \mathrm{SCE}(\boldsymbol{W})}{\partial w_k} = 0$$

- Gradiente descendiente por lotes (BGD)
  - $w_k = w_k - 2\alpha \sum_{r=0}^{e-1} \left(w_k x_{r,k} - y_r\right) \cdot x_{r,k}$
- Gradiente descendiente estocástico (SGD)
  - $w_k = w_k - 2\alpha \left(w_k x_{r,k} - y_r\right) \cdot x_{r,k}$

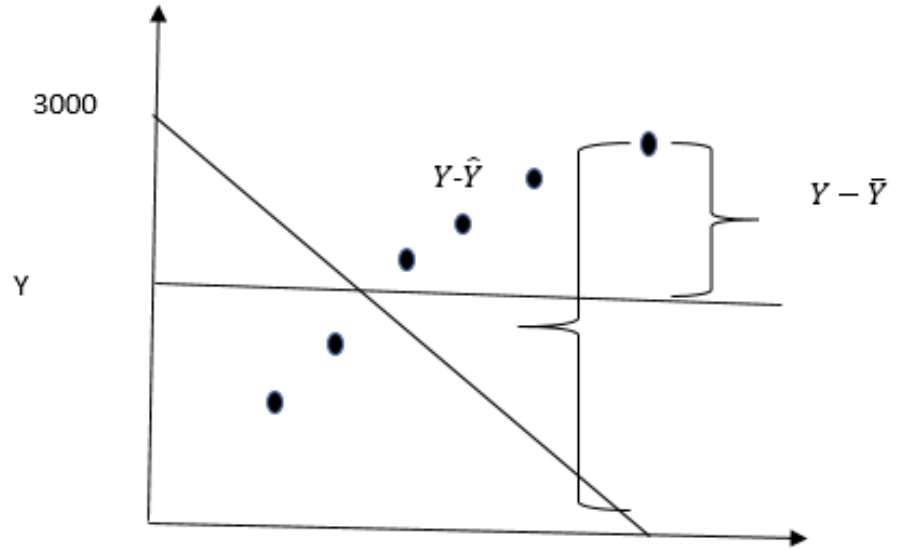
Aprendizaje supervisado

Coeficiente de determinación R^2

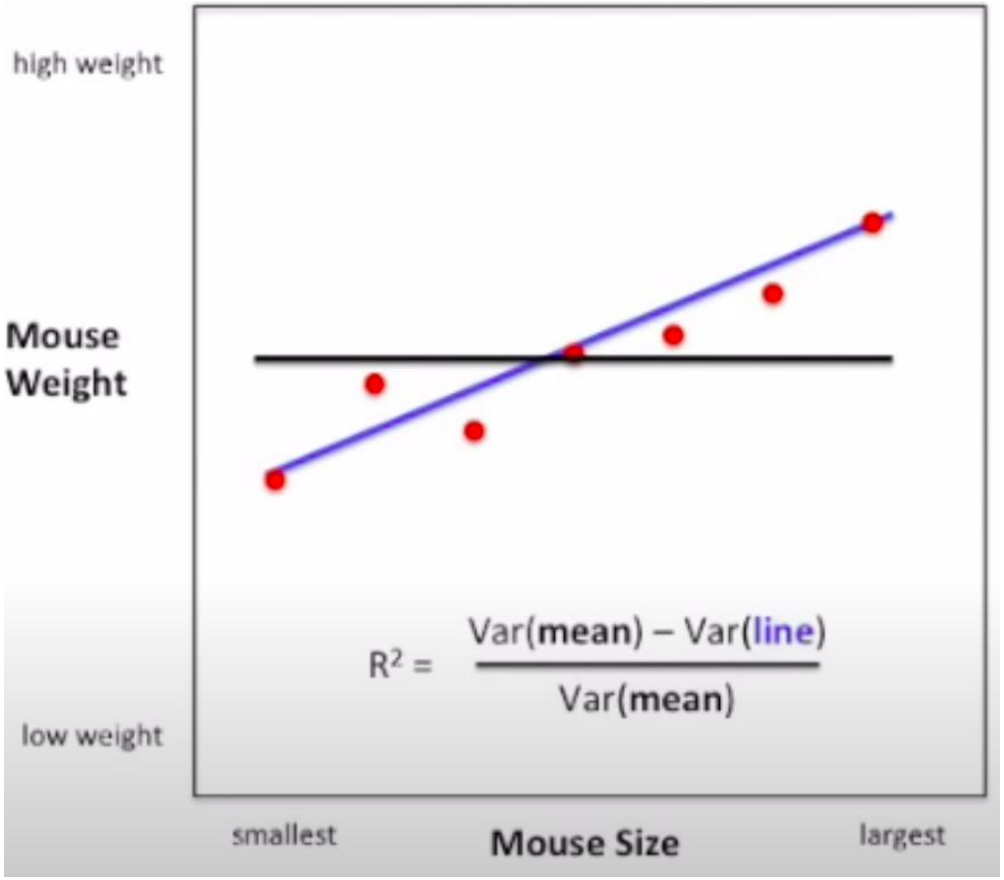
- El coeficiente de determinación, también llamado R cuadrado, refleja que tan bueno es el ajuste de un modelo con respecto a la variable que pretender explicar
- Es la proporción de la varianza total de la variable explicada por la regresión
- Este coeficiente toma valores entre 0 y 1, cuanto más cerca de 1 está, mejor será el ajuste del modelo
- De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será

Coeficiente de determinación R^2

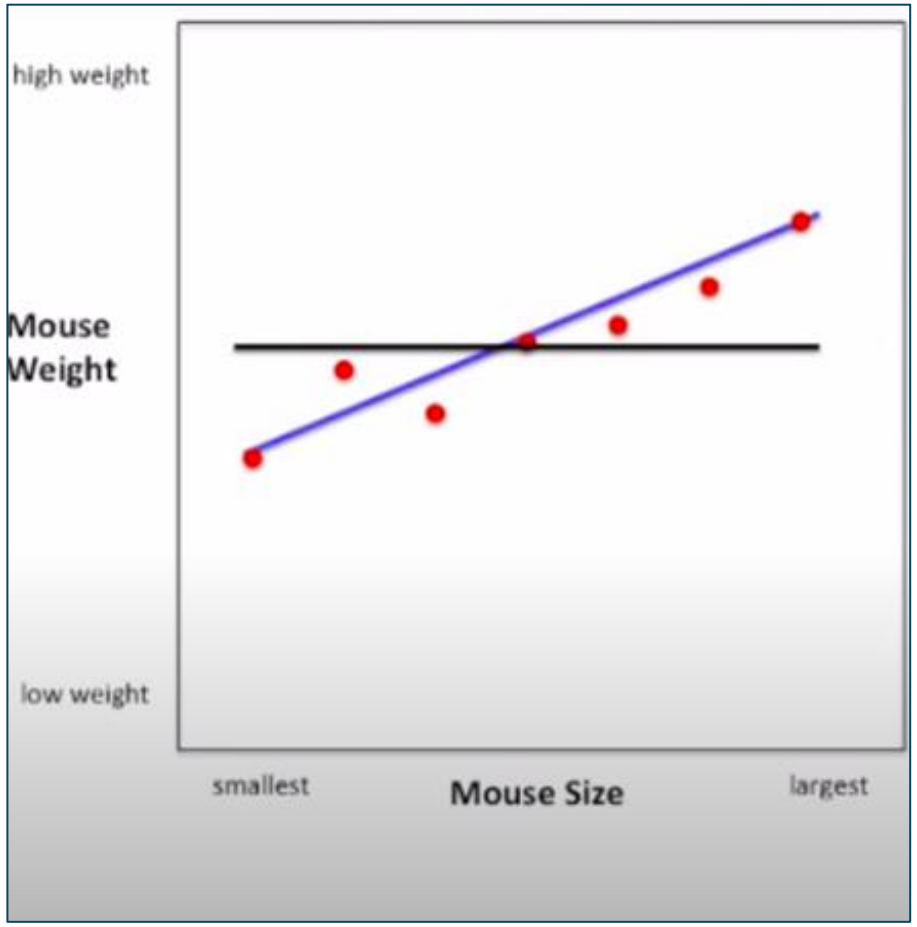
- El valor del coeficiente también puede tomar valores negativos cuando el modelo creado es arbitrariamente incorrecto
- Esto se manifiesta con líneas que no siguen la tendencia de los datos



Coeficiente de determinación R^2



Ejemplo 1. Coeficiente de determinación R^2



$$\text{Var(mean)} = 32$$

$$\text{Var(line)} = 6$$

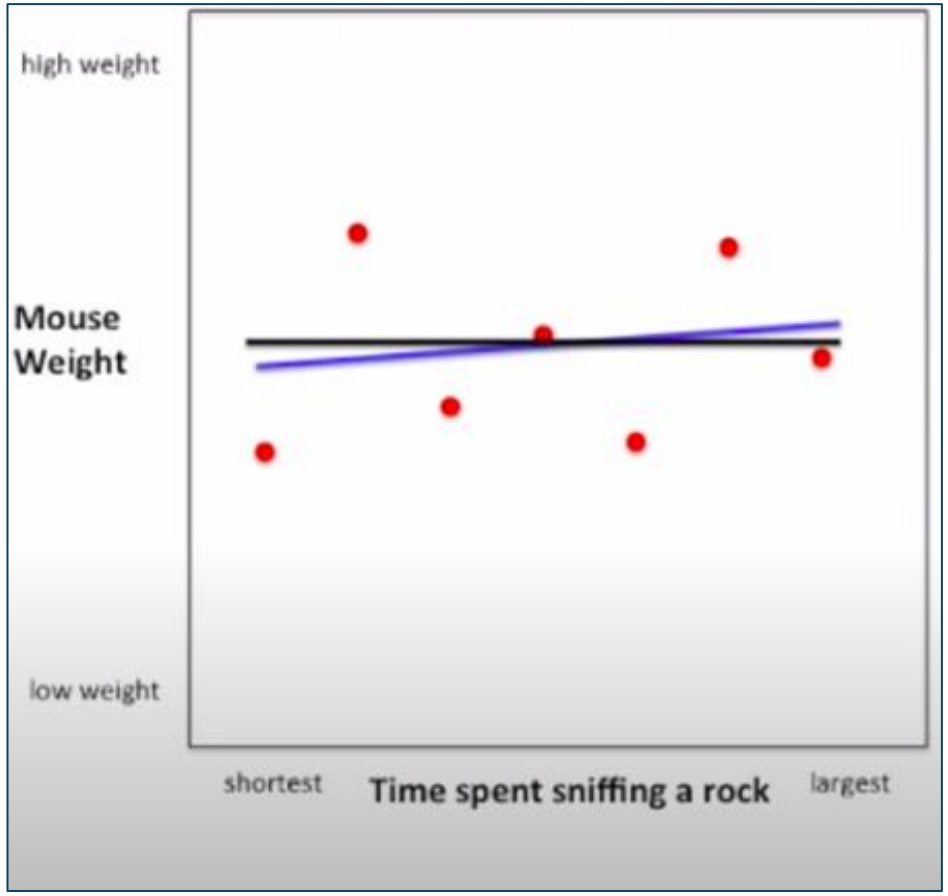
$$R^2 = \frac{\text{Var(mean)} - \text{Var(line)}}{\text{Var(mean)}}$$

$$R^2 = \frac{32 - 6}{32}$$

$$R^2 = \frac{26}{32} = 0.81 = 81\%$$

- Existe un 81% menos de variación entre la línea ajustada y el promedio
- La variable independiente y la salida tienen una alta correlación

Ejemplo 2. Coeficiente de determinación R^2



$$\text{Var(mean)} = 32$$

$$\text{Var(line)} = 30$$

$$R^2 = \frac{\text{Var(mean)} - \text{Var(line)}}{\text{Var(mean)}}$$

$$R^2 = \frac{32 - 30}{32}$$

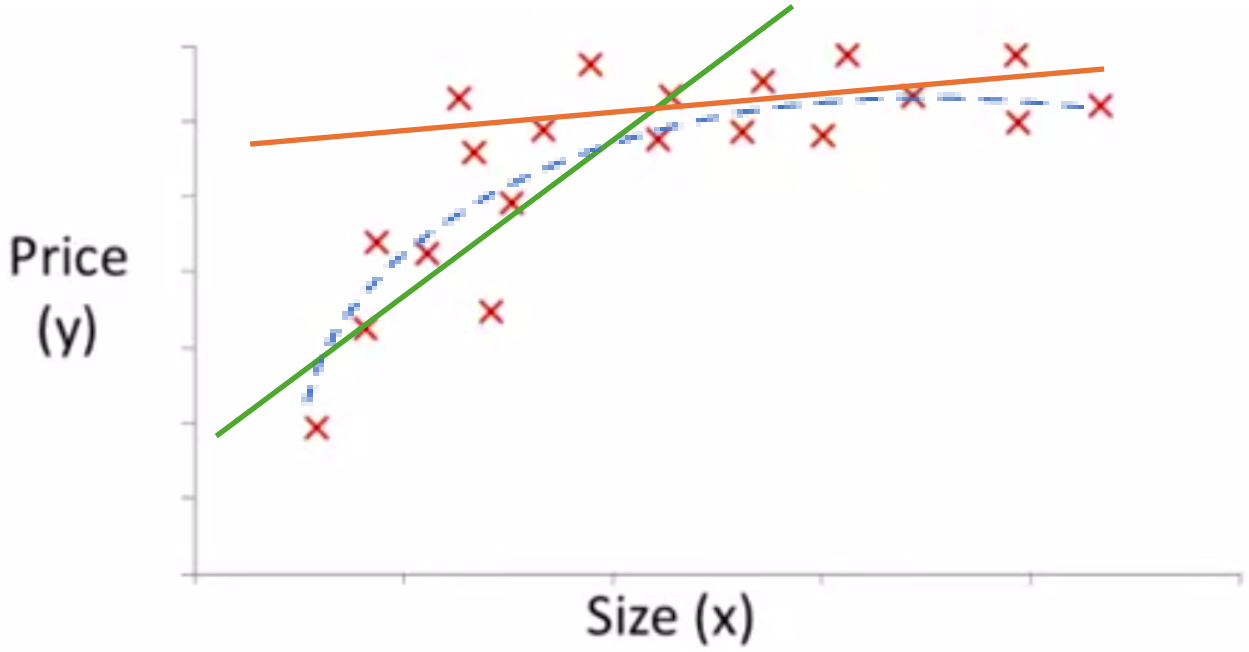
$$R^2 = \frac{2}{32} = 0.06 = 6\%$$

- Sólo hay un 6% menos de variación entre la línea ajustada y el promedio
- La variable independiente y la salida no estás correlacionadas

Limitaciones de la regresión lineal

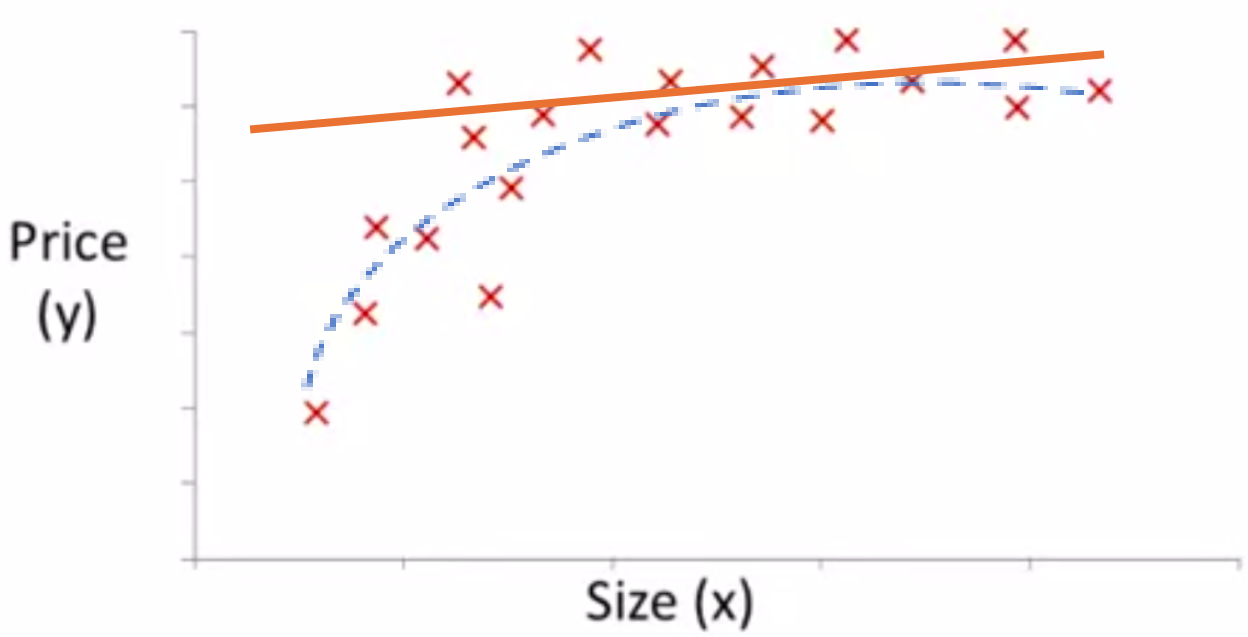
- Los modelos creados mediante regresión vistos hasta ahora presentan un buen desempeño cuando los datos siguen una tendencia lineal
- Sin embargo, pueden existir datos que sigan una tendencia no lineal
- Con este tipo de datos los modelos de regresión lineal obtendrán un bajo desempeño

Datos con tendencia no lineal



Regresión polinomial

- En la regresión polinomial se puede modificar el grado del polinomio para ajustarse a distintas tendencias de los datos



$$\hat{y} = w_0 + w_1 x$$

$$\hat{y} = w_0 + w_1 x + w_2 x^2$$

Escalamiento de los datos

- Algo a considerar en la regresión polinomial es que al elevar a diferentes potencias las variables, sus valores crecerán demasiado
- Esto puede traer como consecuencia que rebasen la capacidad de precisión que se puede manejar en la computadora
- Por ejemplo:
  - x = 1,000
  - X^2 = 1,000,000
  - X^3 = 1,000,000,000
- Además, algunos algoritmos de aprendizaje automático como regresión y redes neuronales muestran sesgo hacia las variables con valores muy grandes

$$w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Escalamiento de datos

- El escalamiento de los datos nos permite establecer un mismo rango de valores para las diferentes las variables
- De esta forma se evita que existan variables con valores muy grandes y otras con valores muy pequeños
- Existen varias técnicas de escalamiento de datos, algunos utilizan la media como medida base para definir el rango, otros utilizan los valores mínimos y máximos para establecer un umbral

Regresión polinomial

- La regresión polinomial permite generar modelos que se ajusten a datos con tendencia no lineal
- En particular esta regresión ayuda cuando se aprecia una tendencia curvilínea entre las variables y el dato a predecir
- Es importante aclarar que la regresión polinomial, en un sentido estricto, se sigue considerando un problema de estimación estadística lineal
- Lo anterior se debe a que la función de regresión sigue siendo lineal a los parámetros (pesos) que se desean estimar a partir de los datos
- Es por eso que la regresión polinomial se considera un caso especial de la regresión lineal multivariable

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_n x^n$$