

Regresión lineal

- Mínimos Cuadrados Ordinarios (OLS)

- Método analítico

$$\frac{\partial \text{SCE}(\mathbf{W})}{\partial w_k} = 0$$

- Gradiente descendiente por lotes (BGD)

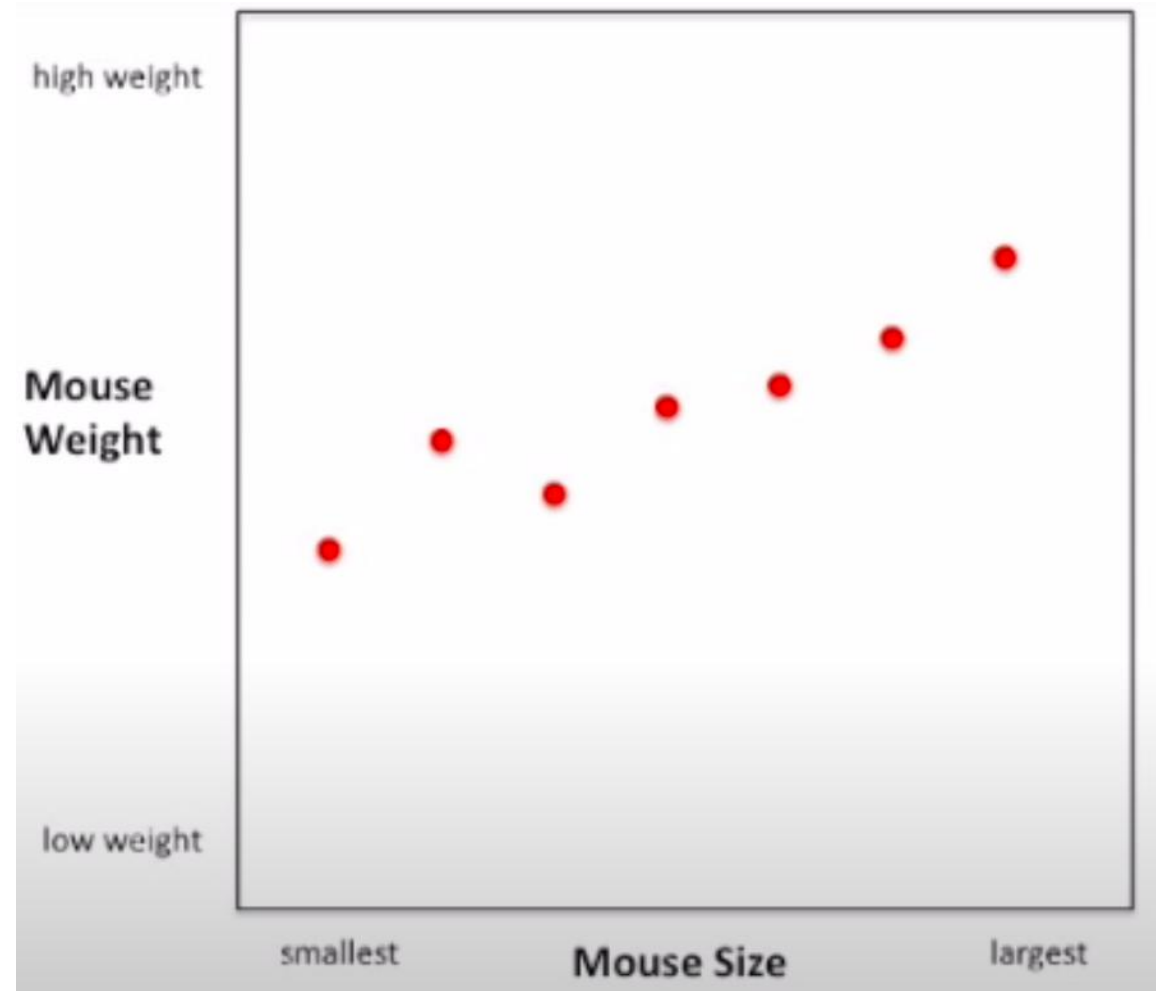
- $w_k = w_k - 2\alpha \sum_{r=0}^{e-1} (w_k x_{r,k} - y_r) \cdot x_{r,k}$

- Gradiente descendiente estocástico (SGD)

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

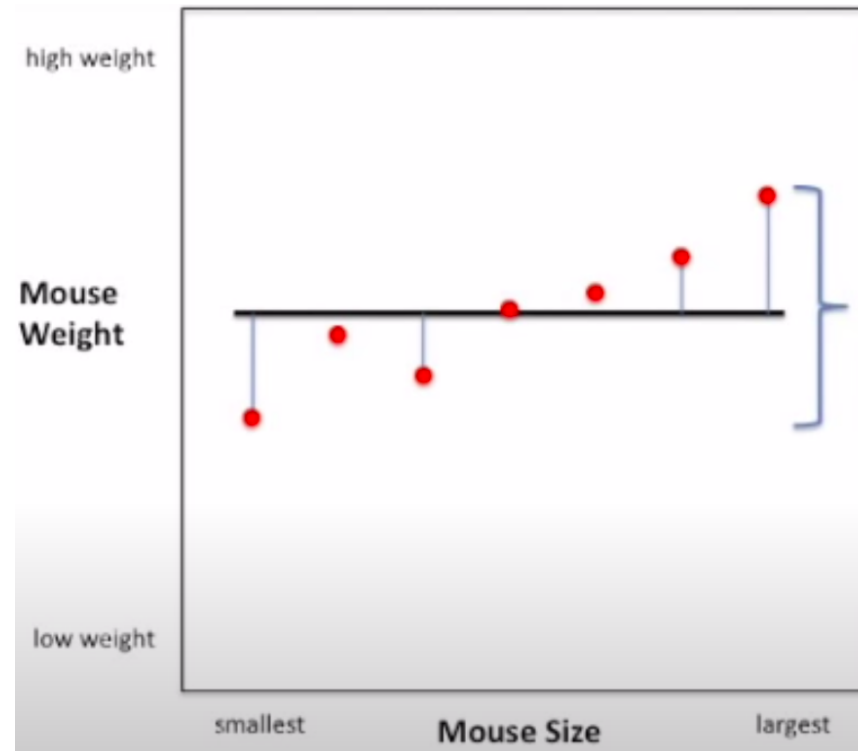
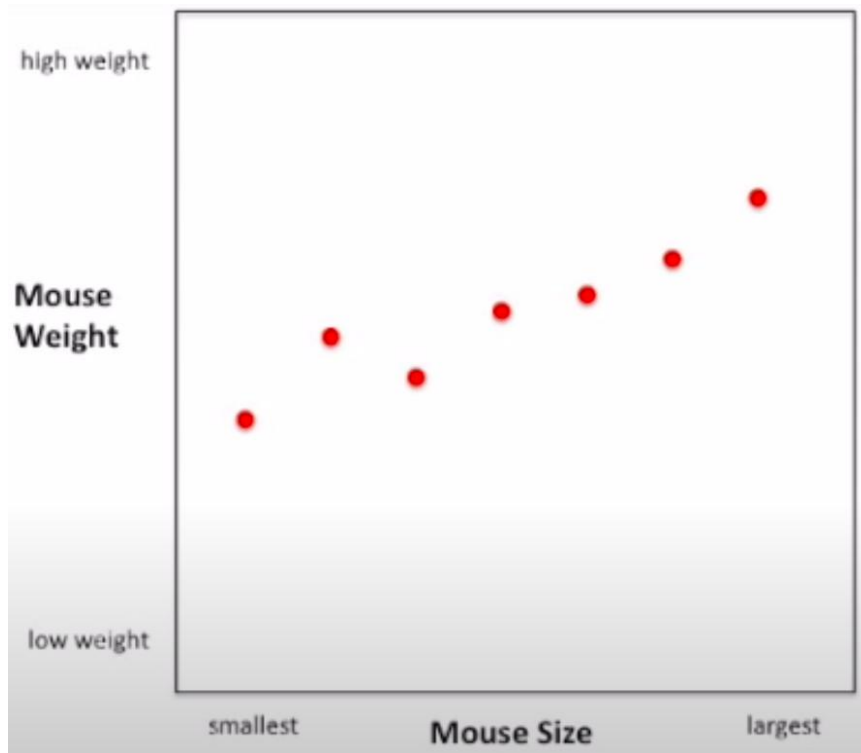
Aprendizaje supervisado

- OLS, BGD y SGD
- ¿Existe una forma más simple de hacer regresión?

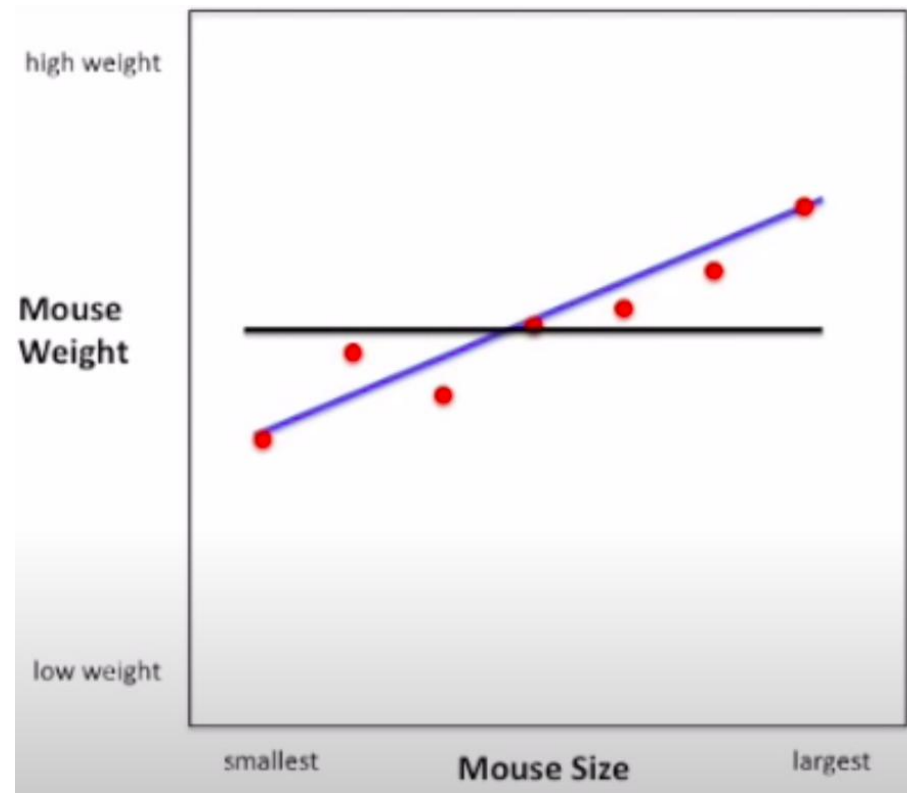


Diferencias usando el promedio de los valores

La forma más simple de regresión es mediante el promedio $\frac{\sum_{i=1}^m y_i}{m}$



Promedio vs línea ajustada

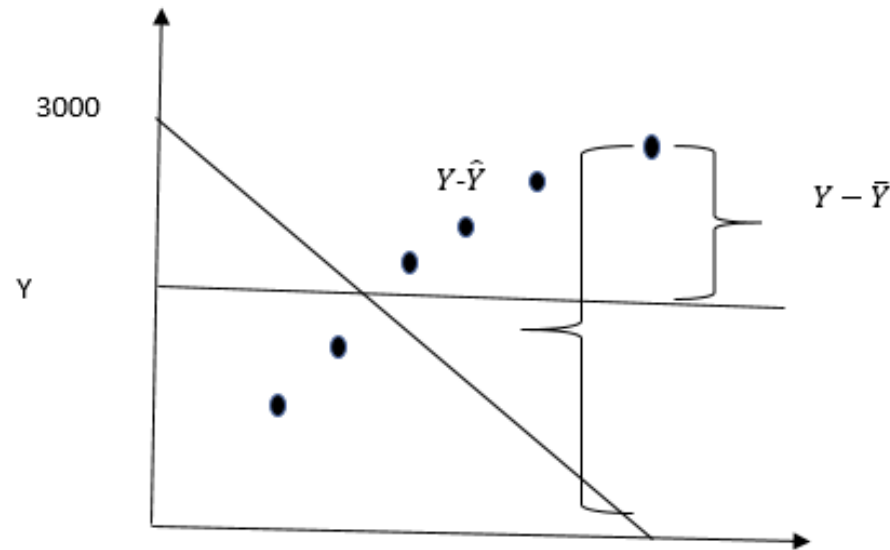


Coeficiente de determinación R^2

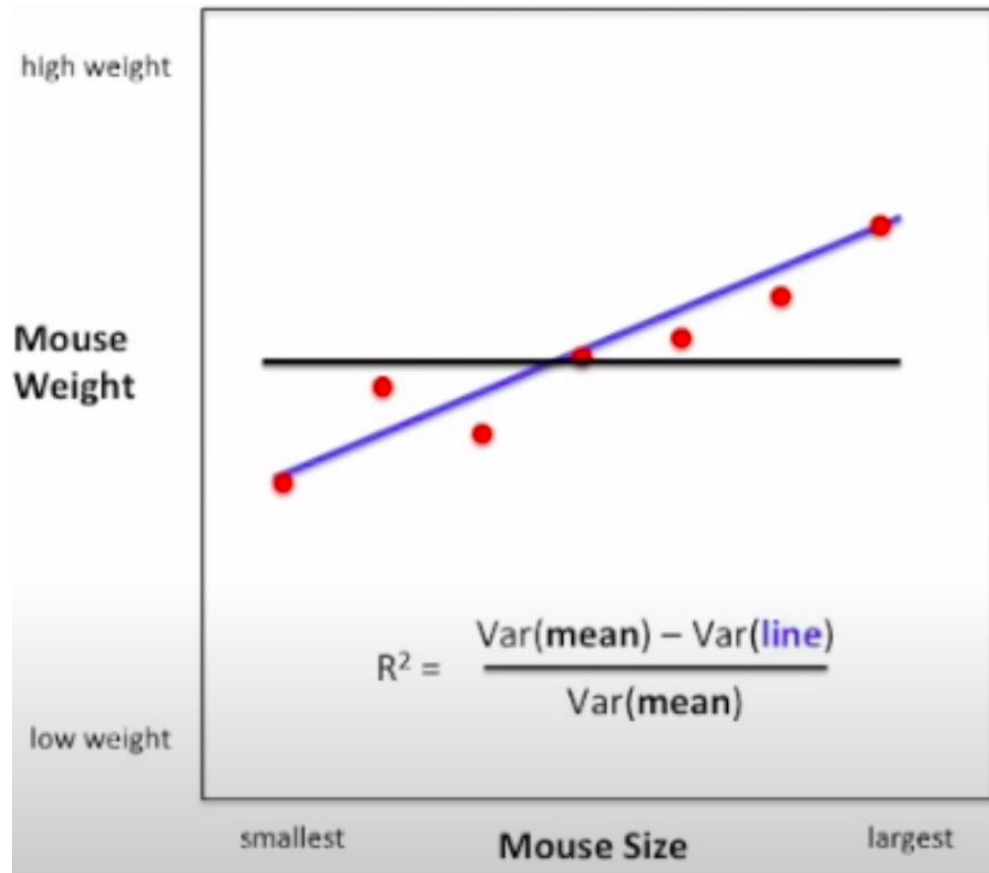
- El coeficiente de determinación, también llamado R cuadrado, refleja que tan bueno es el ajuste de un modelo con respecto a la variable que pretender explicar
- Es la proporción de la varianza total de la variable explicada por la regresión
- Este coeficiente toma valores entre 0 y 1, cuanto más cerca 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

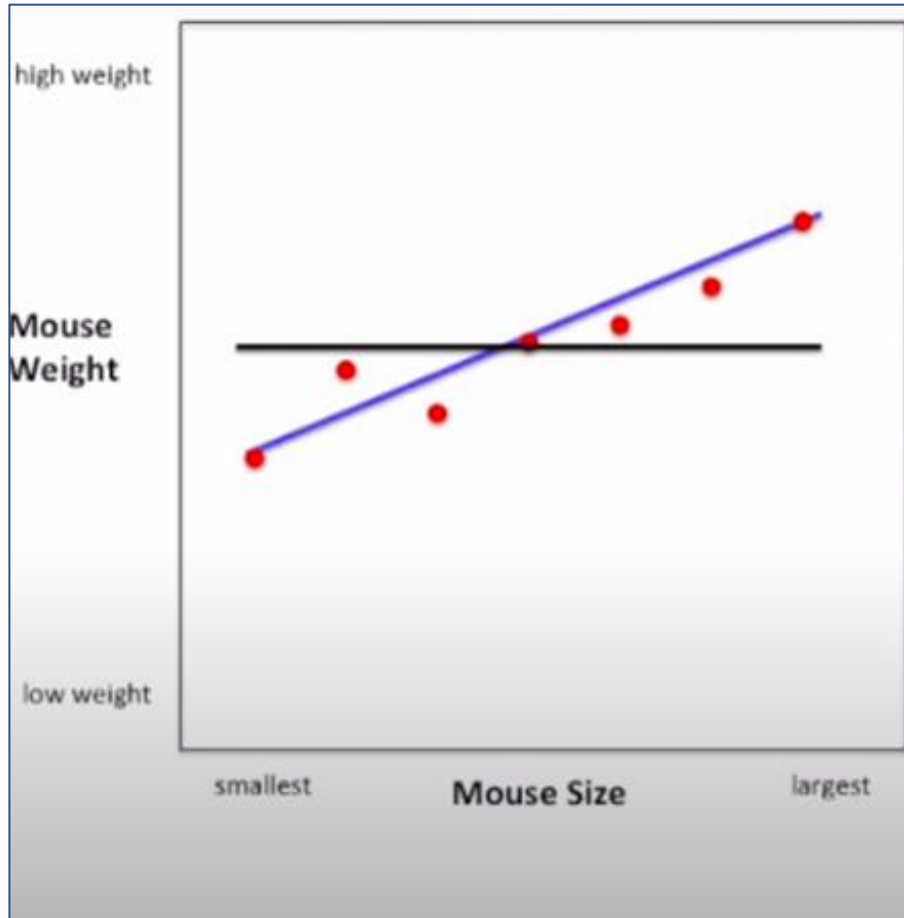
- El valor del coeficiente también puede tomar valores negativos cuando el modelo creado es arbitrariamente incorrecto
- Esto se manifiesta con líneas que no siguen la tendencia de los datos



Coeficiente de determinación R^2



Ejemplo 1. Coeficiente de determinación R^2



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

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

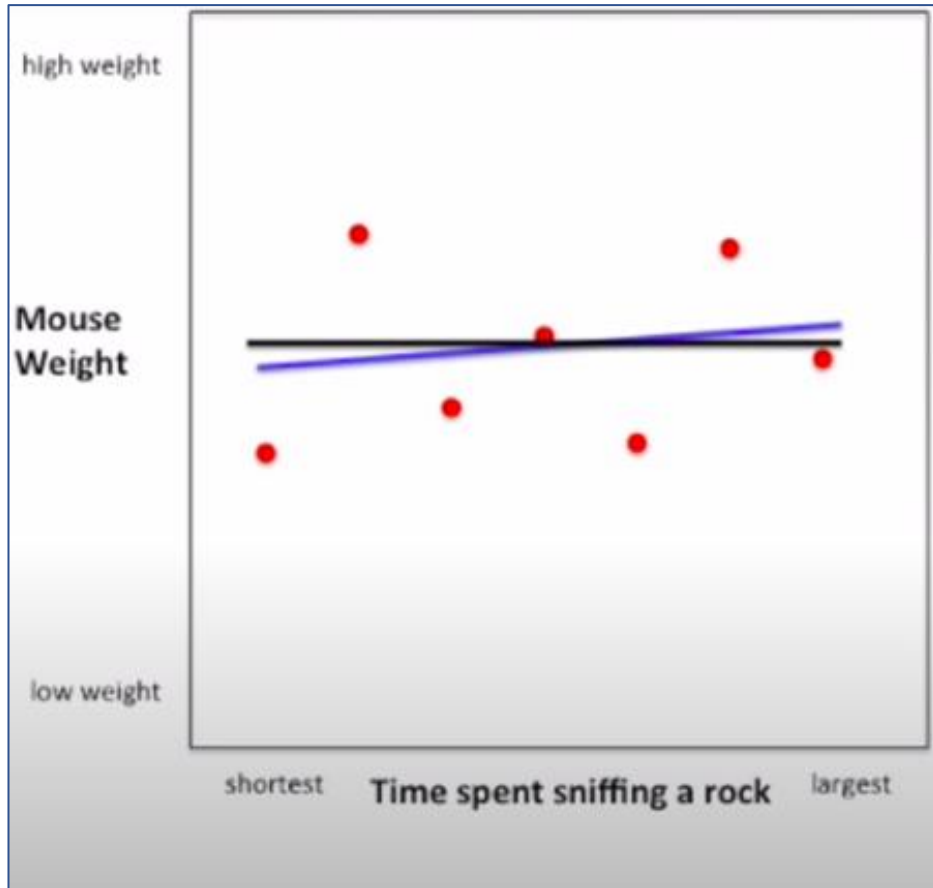
$$R^2 = \frac{\text{Var}(\text{mean}) - \text{Var}(\text{line})}{\text{Var}(\text{mean})}$$

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

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

- 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



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

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

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

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

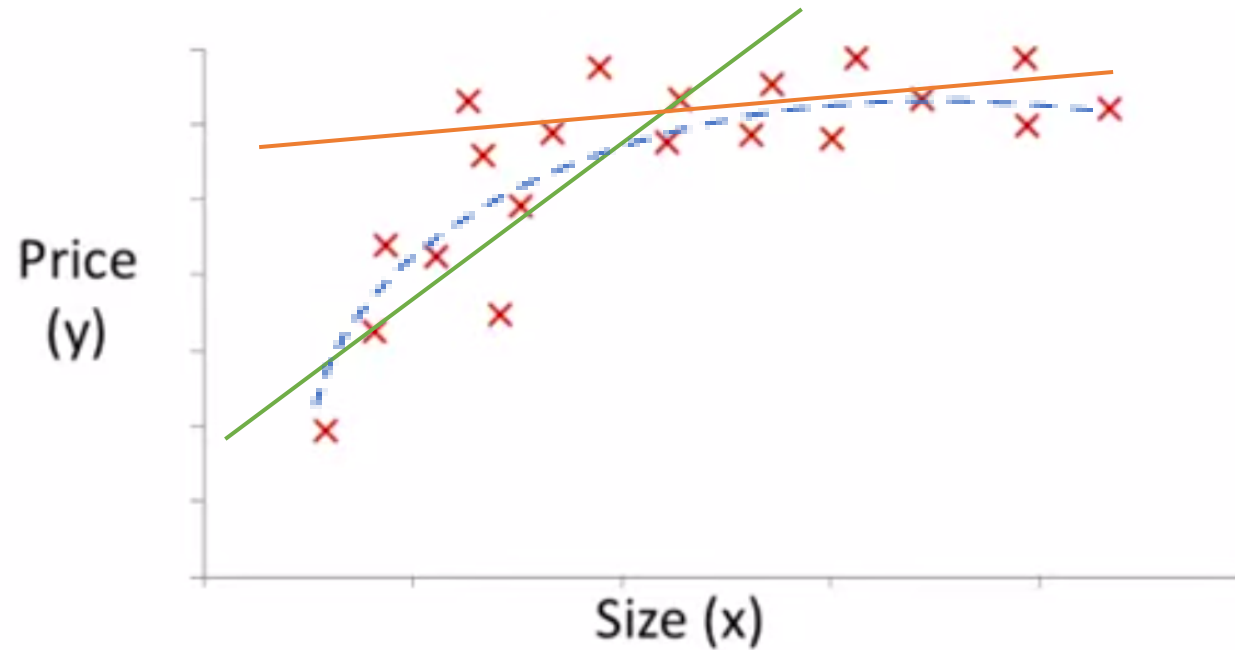
$$R^2 = \frac{2}{32} = 0.06 = 6\%$$

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

Limitaciones de la regresión lineal

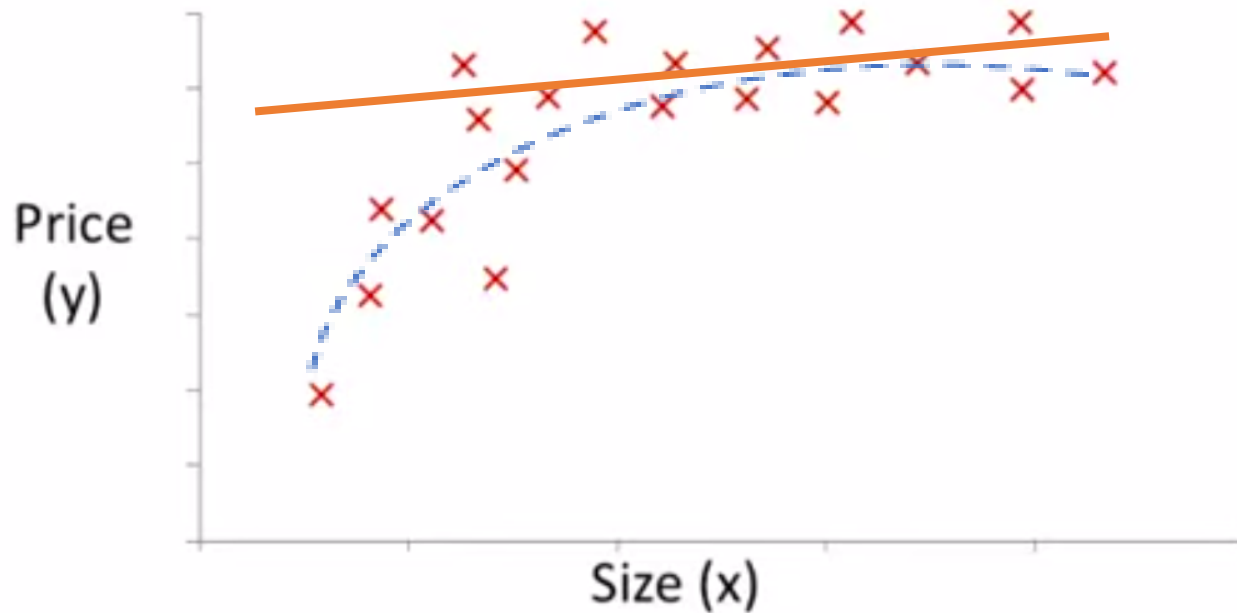
- Los modelos creados mediante regresión vistos hasta ahora presentan un buen desempeño cuando los datos siguen una tendencia lineal
- Sin embargo, pueden existir datos que sigan una tendencia no lineal
- Con este tipo de datos los modelos de regresión lineal obtendrán un bajo desempeño

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



$$\hat{y} = w_0 + w_1x$$

$$\hat{y} = w_0 + w_1x + w_2x^2$$

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

$$w_0 + w_1x + w_2x^2 + w_3x^3$$

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

$$\hat{y} = w_0 + w_1x + w_2x^2 + w_3x^3 + \dots + w_nx^n$$