Introducción al modelado

¿Qué es modelar datos?

¿Qué es modelar datos?

Un modelo es una explicitación simplificada de algunas relaciones potencialmente existentes entre las variables disponibles de los datos.

¿Qué es modelar datos?

Un modelo es una explicitación simplificada de algunas relaciones potencialmente existentes entre las variables disponibles de los datos.

Representación → Evaluación → Optimización

Representación: Buscamos una manera de modelar una variable en función de otra (u otras).

Evaluación: Definimos alguna medida que nos permita evaluar que tan bien ajusta el modelo.

Optimización: Dada la representación y la evaluación definimos un procedimiento efectivo que pueda hallar la determinación óptima de los parámetros del modelo.

Supervisados vs. No Supervisados

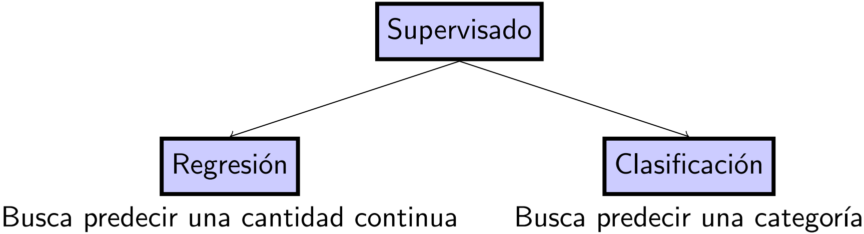
Supervizados

- Disponemos de un conjunto de datos de entrada y salida etiquetados.
- Aprenden a hacer predicciones basados en patrones de los datos del modelo.
- Ej: Regresión lineal, árboles de decisión, regresión logística, redes neuronales, etc.

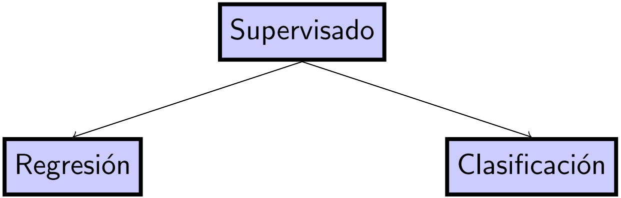
No Supervisados

- No disponemos de etiquetas o resultados deseados en el conjunto de datos.
- Buscan patrones y estructuras ocultas en los datos sin guía previa.
- Ej: Métodos de clustering (K-means, jerárquico), PCA, etc.

Regresión vs. Clasificación



Regresión vs. Clasificación



Busca predecir una cantidad continua

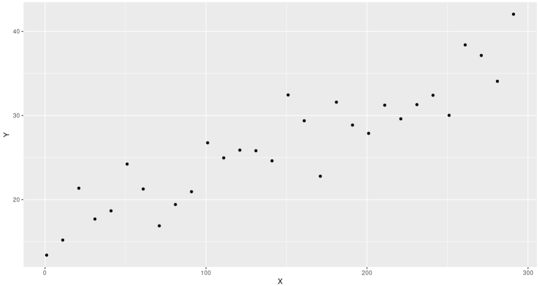
Busca predecir una categoría

Ejemplos:

Regresión: Predecir el *peso de los pingüinos* en función de la longitud del pico y la longitud de las aletas.

Clasificación: Predecir la *especie de pingüino* en función de la longitud del pico y la longitud de las aletas.

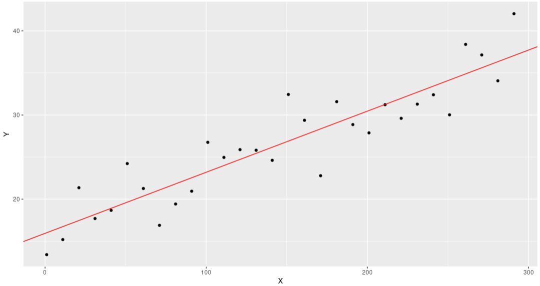
Modelo de Regresión Lineal Simple



Modelo matemático: $Y = \beta_0 + \beta_1 X$

- β_0 es la ordenada al origen (o intercept)
- β_1 es la pendiente

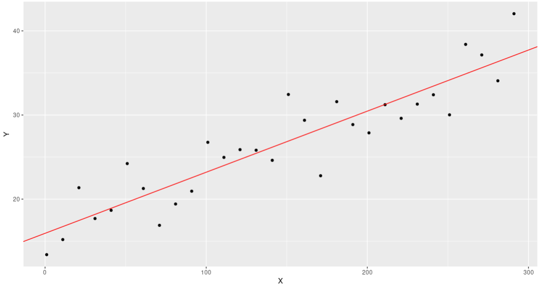
Modelo de Regresión Lineal Simple



Modelo matemático: $Y = \beta_0 + \beta_1 X$

- β_0 es la ordenada al origen (o intercept)
- β_1 es la pendiente

Modelo de Regresión Lineal Simple



Modelo matemático: $Y = \beta_0 + \beta_1 X$

- β_0 es la ordenada al origen (o intercept)
- β_1 es la pendiente

Modelo de regresión lineal: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

- x_i es la variable predictora
- y_i es la variable dependiente
- ϵ_i es un error aleatorio (variación de Y no explicada por X)
- β_0 y β_1 son los parámetros del modelo

Modelo de Regresión Lineal Simple

Estimamos los parámetros β_0 y β_1 usando los datos:

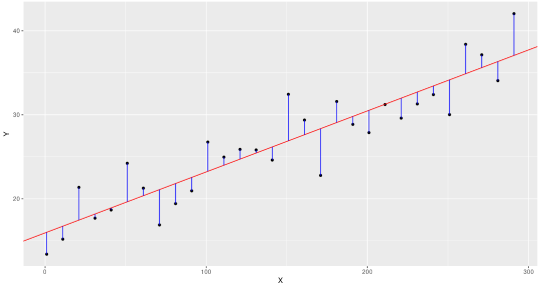
$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Y con los parámetros estimamos la variable dependiente

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

¿Cómo encuentro la mejor recta?

Método de Mínimos Cuadrados



Residuo: Es la diferencia entre el valor observado y el predicho: $y_i - \widehat{y}_i$

Definimos la suma de los residuos al cuadrado (RSS): $RSS = (y_1 - \widehat{y}_1)^2 + \cdots + (y_n - \widehat{y}_n)^2$

La mejor recta va a ser la que minimice:

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Método de mínimos cuadrados

Podemos derivar respecto de los párámetros, igualar a cero para buscar puntos críticos y despejar. Así encontramos β_0 y β_1 para minimizar la RSS.

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

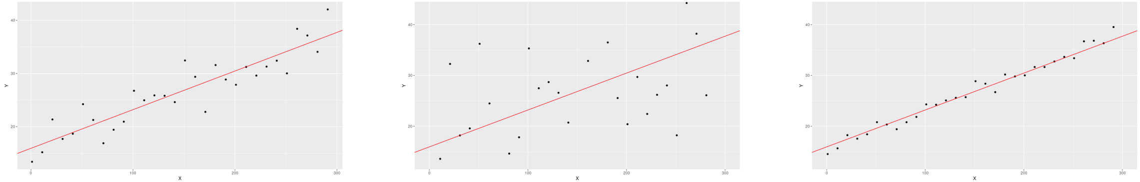
con, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Evaluación del modelo

Error cuadrático medio (MSE) Cuantifica qué tan cerca está un valor predicho del valor real, por lo que se puede usar para cuantificar qué tan cerca está una línea de regresión de un conjunto de puntos.

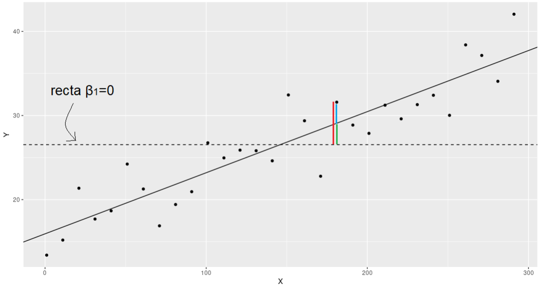
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Varianza del modelo



Si tenemos distintos escenarios para la misma recta, ¿cómo medimos la variabilidad?

Varianza del modelo



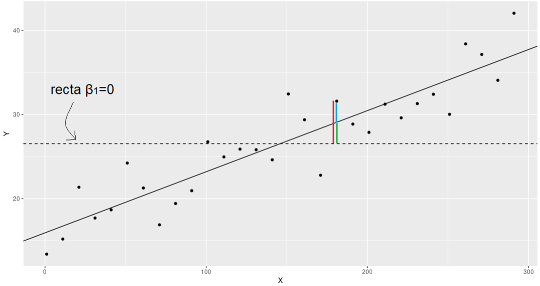
La variabilidad de total del modelo se puede separar en explicada y no explicada.

Variabilidad total $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Variabilidad no explicada $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$

Variabilidad explicada $\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$

Varianza del modelo



La variabilidad de total del modelo se puede separar en explicada y no explicada.

La proporción de la variabilidad de Y explicada por X se puede explicar como:

$$R^2 = \frac{\text{Variabilidad explicada}}{\text{Variabilidad total}}$$

$0 \leq R^2 \leq 1$

A mayor R^2 más cercanos los puntos a la recta y mayor "fuerza" para predecir.