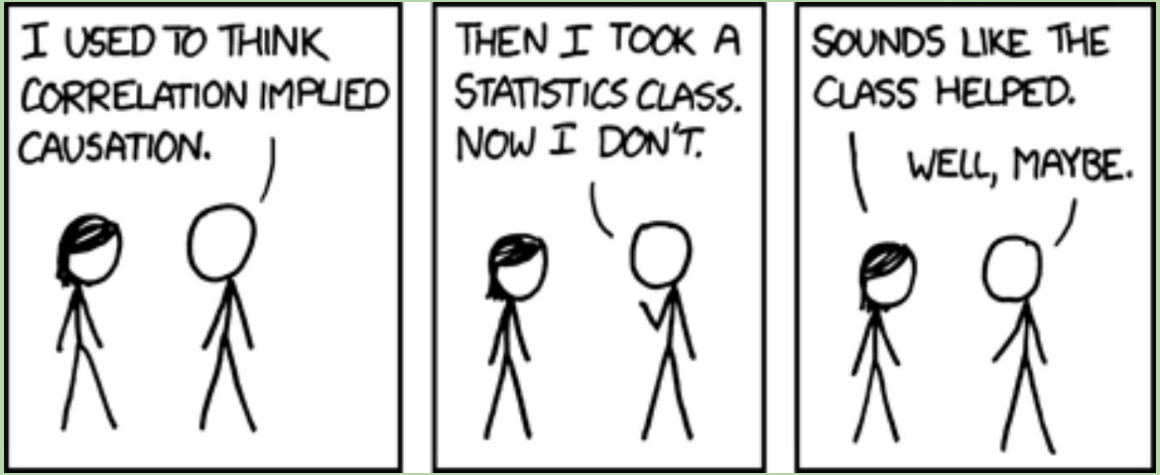
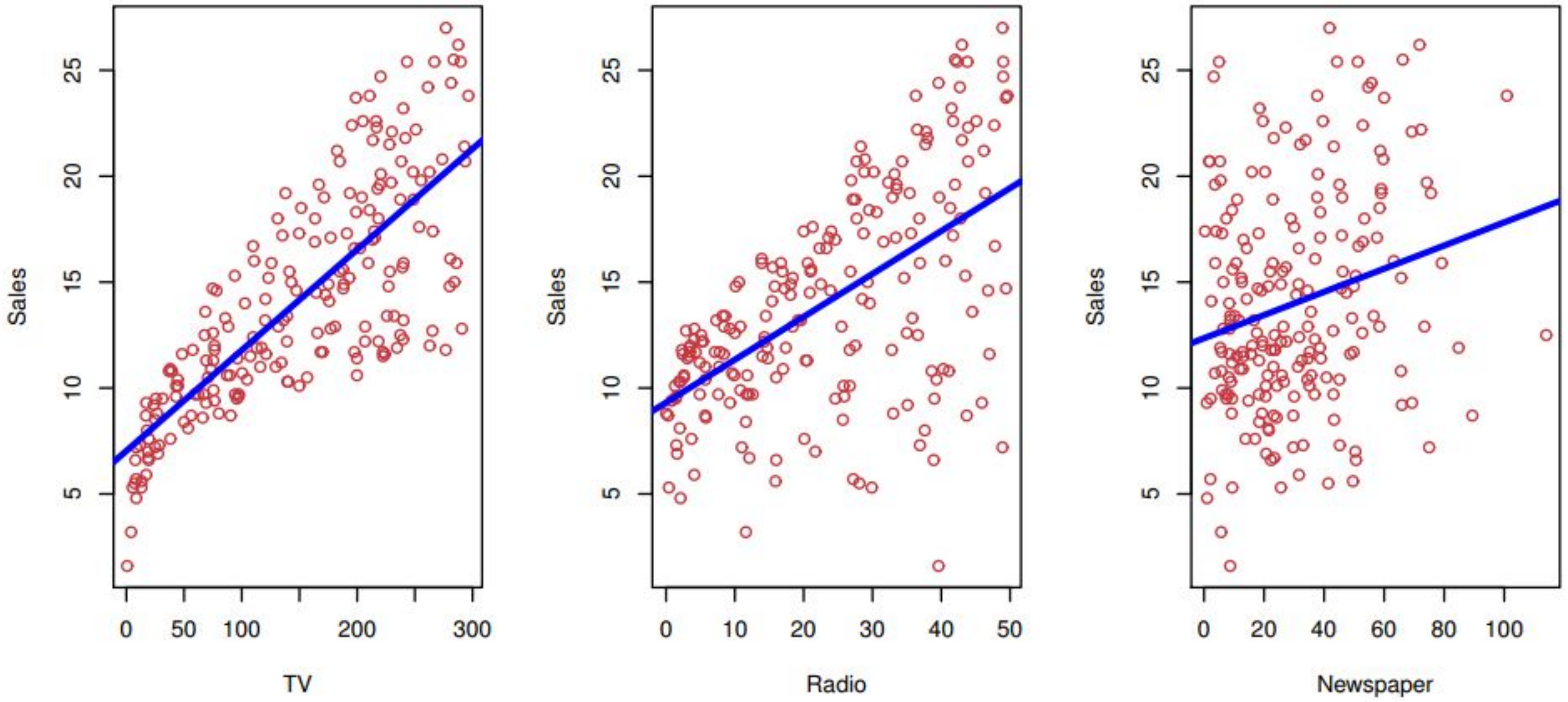
Introducción al curso

Bienvenidxs!



Un ejemplo...



Sales
5
10
15
20
25
0
50
100
200
300
TV
Sales
5
10
15
20
25
0
10
20
30
40
50
Radio
Sales
5
10
15
20
25
0
20
40
60
80
100
Newspaper

Tipos de variables:

Input variable (Variable de entrada): X

X_1: *Presupuesto de la TV (en relación al ejemplo)*
X_2: *Presupuesto de la radio*
X_3: *Presupuesto de los periódicos.*

Los inputs reciben distintos nombres, como ***predictores***, ***variables independientes*** o a veces simplemente variables.

Output variable (variable de salida): Y

Y: *Ventas (en relación al ejemplo)*

La variable de salida suele denominarse ***variable de respuesta*** o ***dependiente***.

Supongamos que observamos una respuesta cuantitativa Y y predictores diferentes, $X_1, X_2, \ldots, X_p$. Suponemos que existe alguna relación entre Y y $X = (X_1, X_2, \ldots, X_p)$, que se puede escribir de forma muy general:

$$Y = f(X) + \epsilon$$

¿Por qué estimar ***f***?

Predicción

$$\hat{y} = \hat{f}(X)$$

$\hat{f}$: representa nuestra estimación para f.

$\hat{y}$: representa la predicción resultante para Y.

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\operatorname{Var}(\epsilon)}_{\text{Irreducible}} \quad ,
\end{aligned}
$$

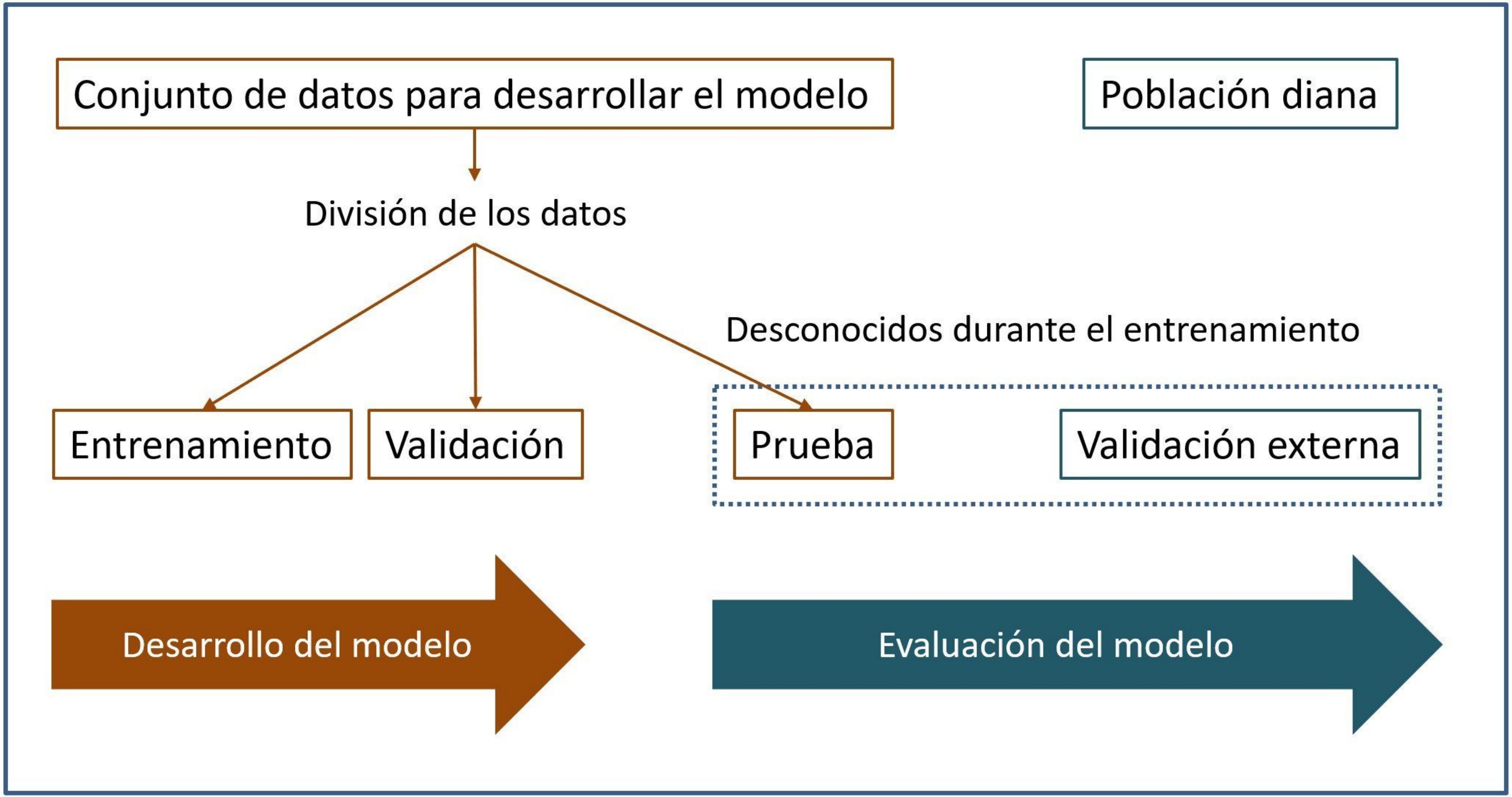
Podemos mejorar potencialmente la precisión de $\hat{f}$ utilizando la técnica de aprendizaje estadístico más adecuada para estimar *f*.

Puede contener variables no medidas que sean útiles para predecir Y. Puede contener también una variación no mensurable.

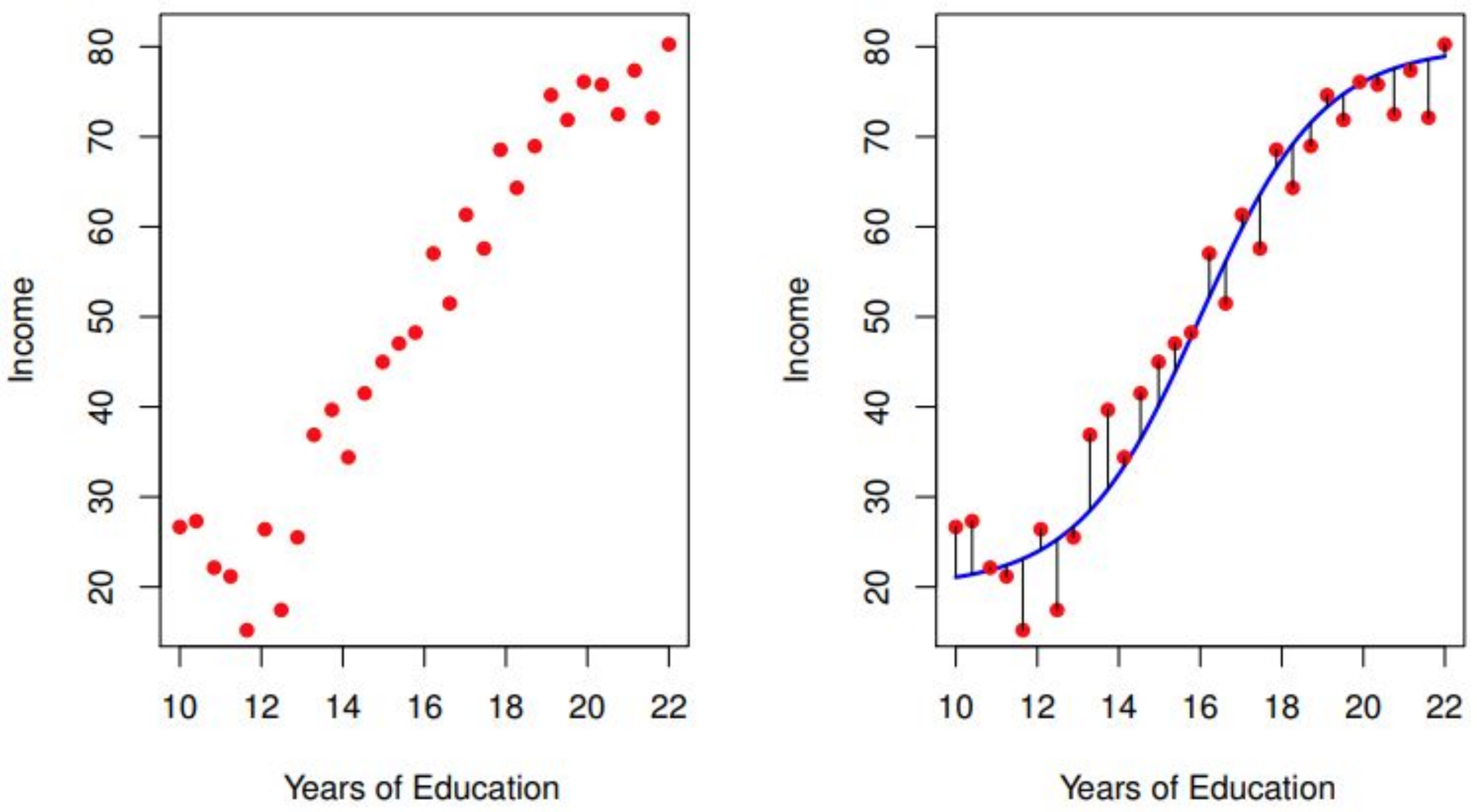
¿Cómo estimamos *f* en el caso de los modelos predictivos?

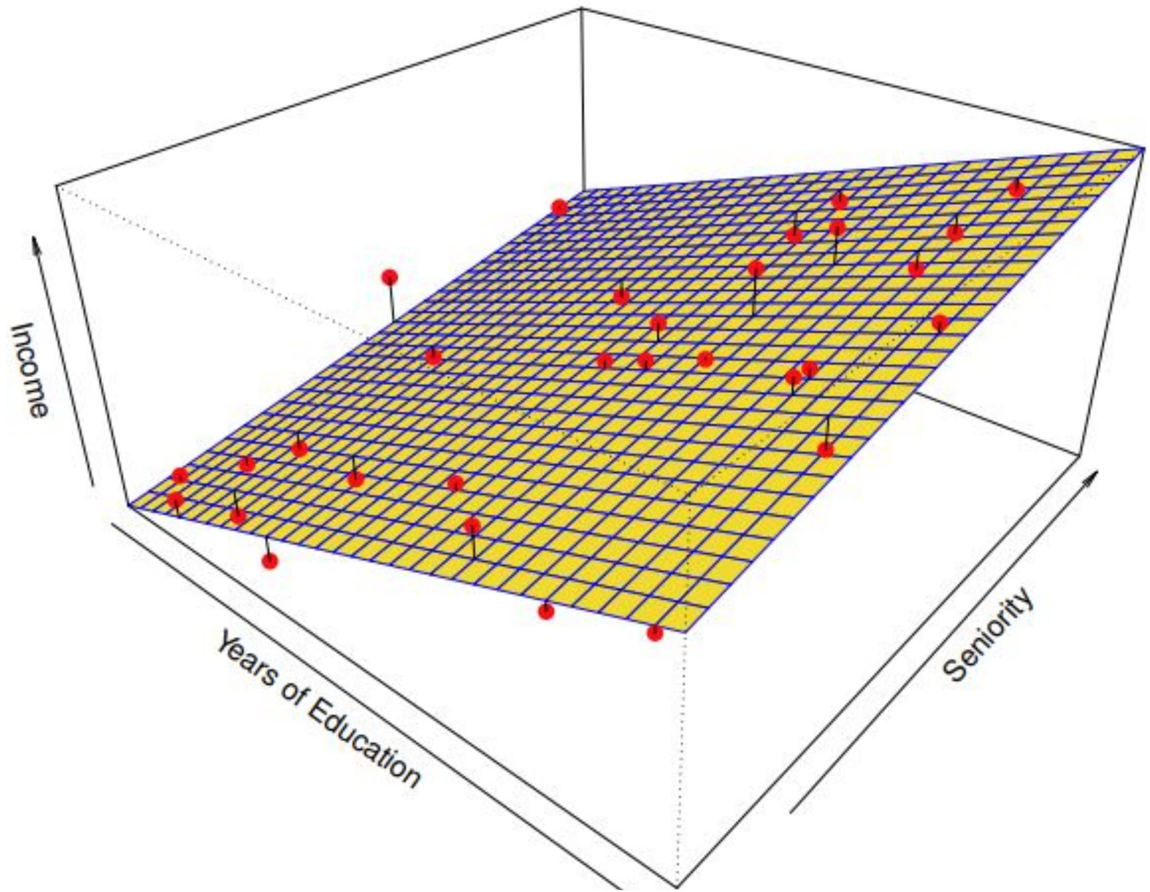
$\hat{f}$ tal que $Y \approx (X)$ para cualquier observación (X, Y).

Datos de entrenamiento


Conjunto de datos para desarrollar el modelo
Población diana
División de los datos
Desconocidos durante el entrenamiento
Entrenamiento
Validación
Prueba
Validación externa
Desarrollo del modelo
Evaluación del modelo

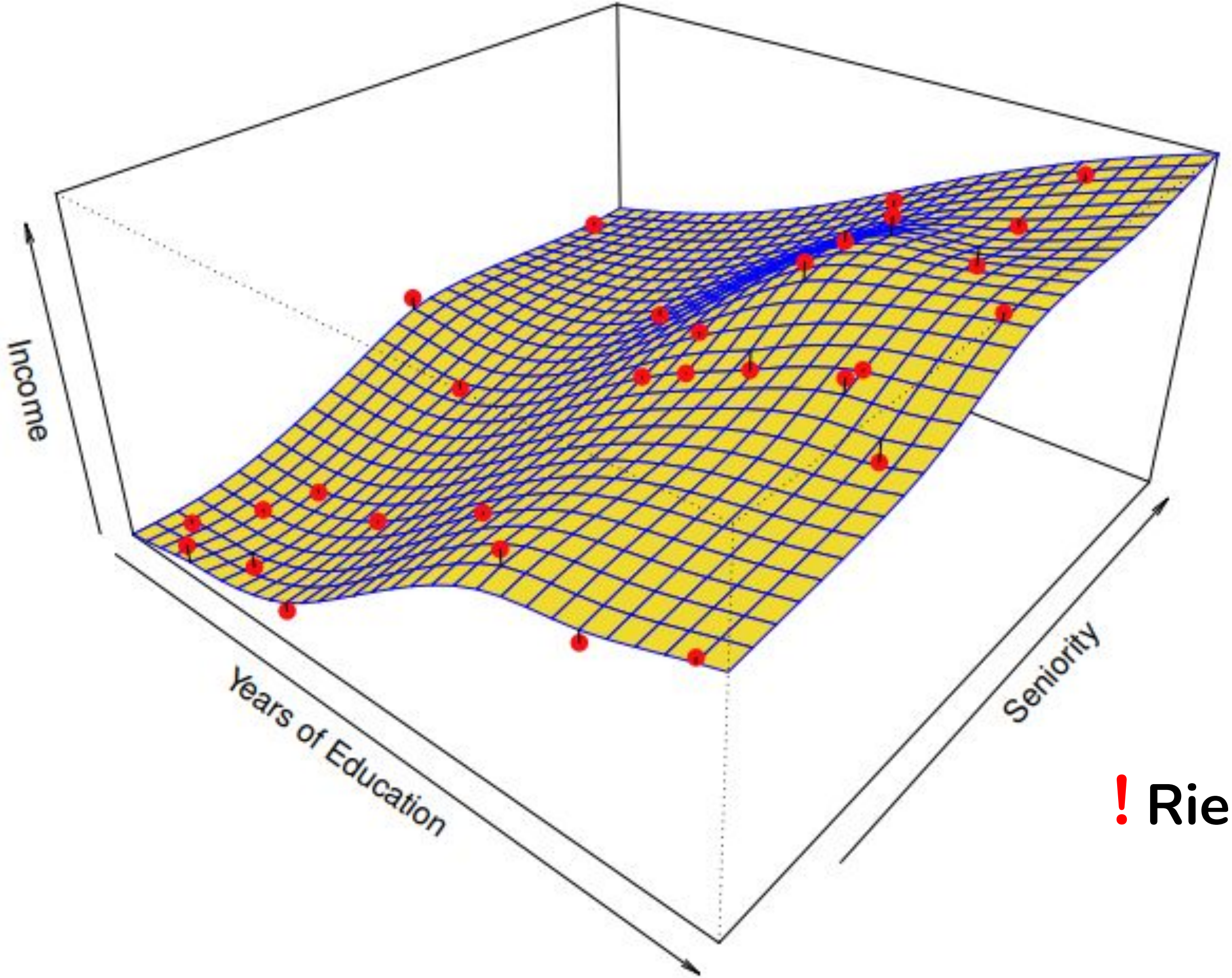
Otro ejemplo...



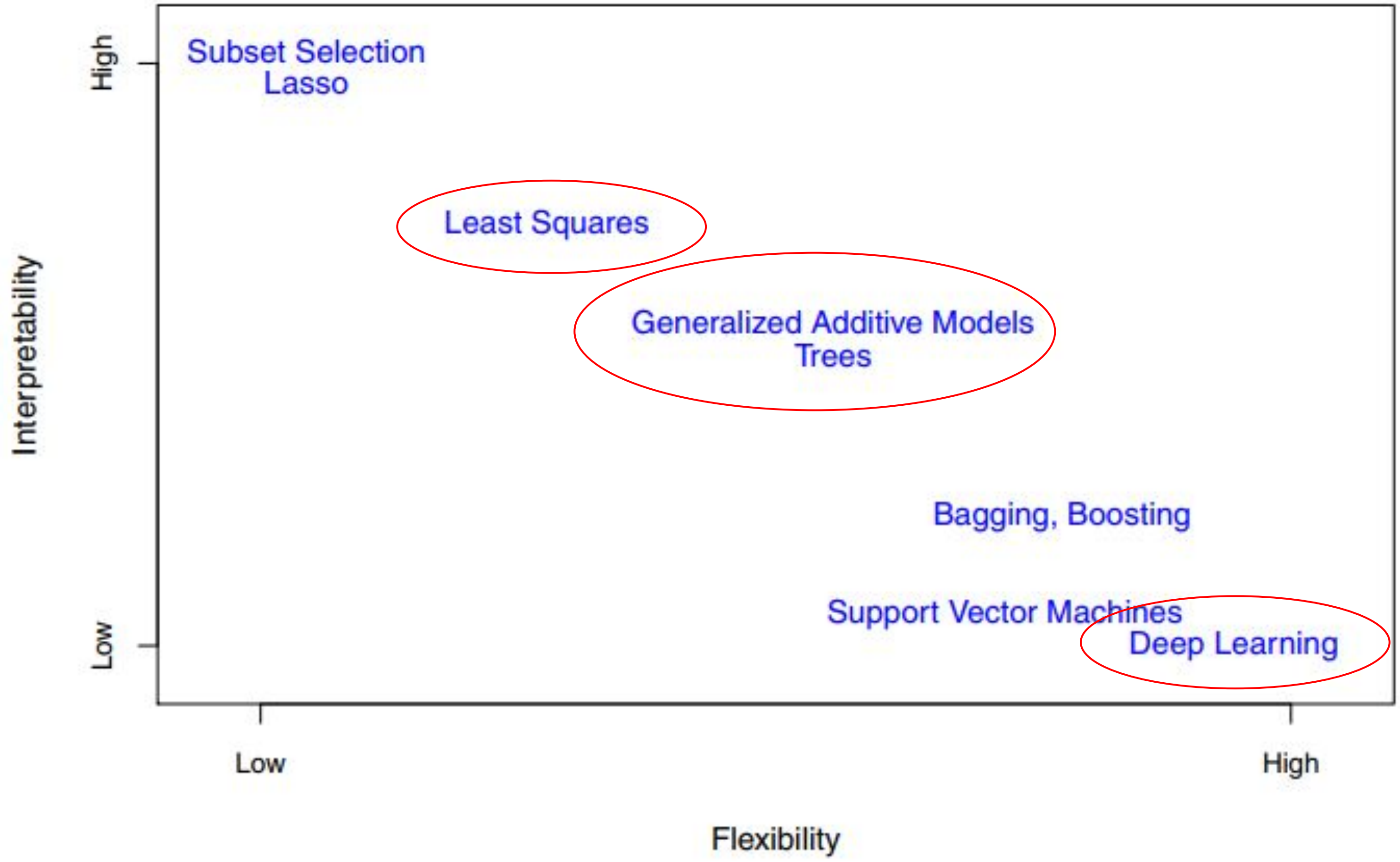


- ❑ **Paramétrico**
- ❑ **Modelo Lineal**
- ❑ **Mínimos cuadrados**

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$



! Riesgo de overfitting



High
Low
Interpretability
Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
Deep Learning
Low
High
Flexibility

Statistical Learning

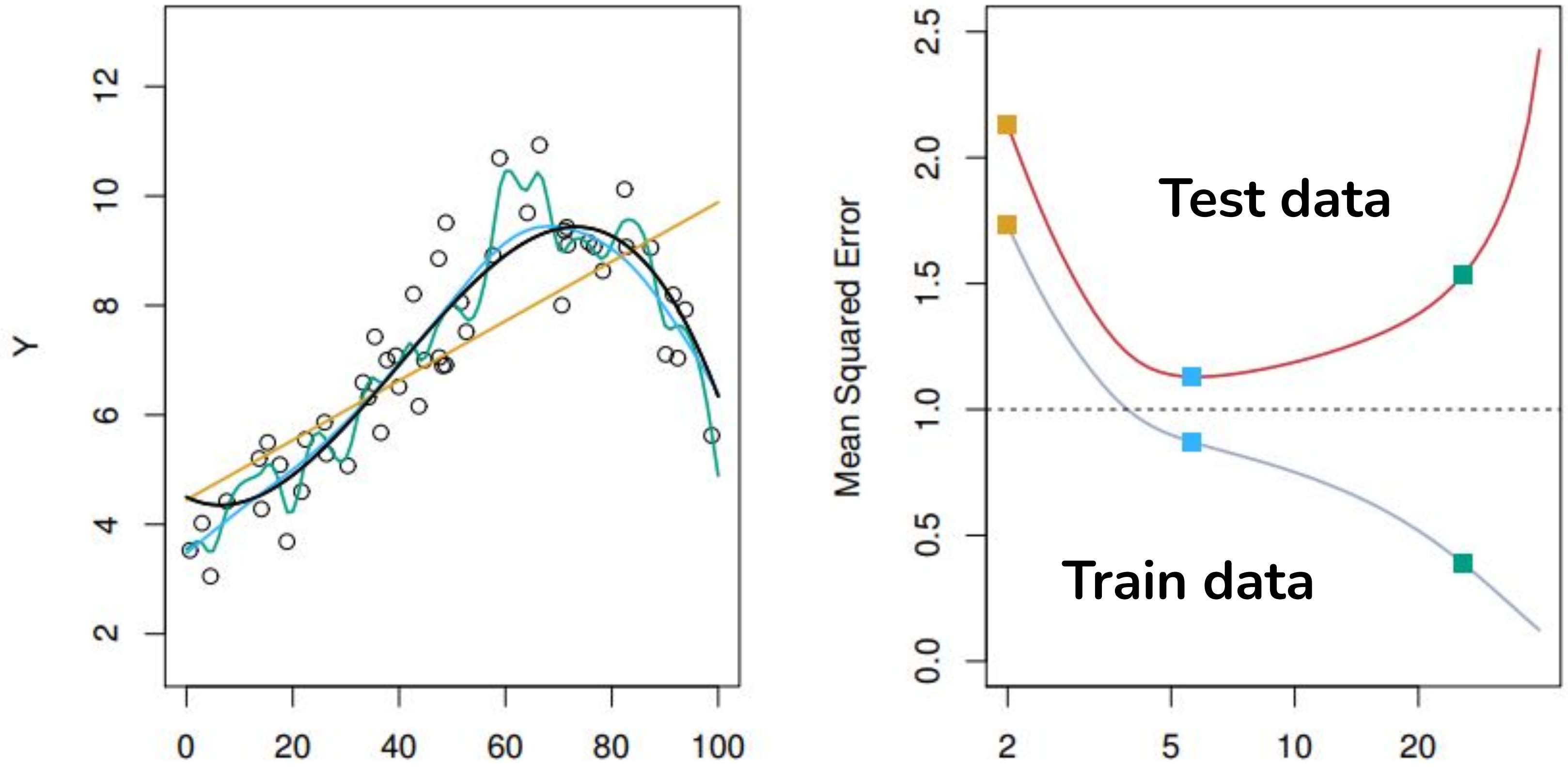
Aprendizaje Supervisado

Datos etiquetados

Aprendizaje No Supervisado

Datos no etiquetados

¿Cómo evaluamos la calidad del ajuste de nuestro modelo?



¿Qué entendemos por varianza y sesgo de un método de aprendizaje estadístico?

- La varianza se refiere a la cantidad en la que $\hat{f}$ cambiaría si la estimamos utilizando un conjunto de datos de entrenamiento diferente. En general, los métodos estadísticos más flexibles tienen mayor varianza.

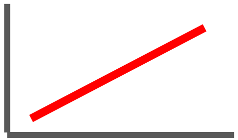
- El sesgo se refiere al error que se introduce al aproximar un problema de la vida real, que puede ser extremadamente complicado, mediante un modelo mucho más sencillo. Por ejemplo, la regresión lineal supone que existe una relación lineal entre Y y $X_1, X_2, \ldots, X_p$. Es poco probable que un problema de la vida real tenga una relación lineal tan sencilla por lo que realizar una regresión lineal dará lugar sin duda a cierto sesgo en la estimación de f.

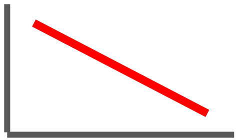
Por regla general, a medida que utilicemos métodos más flexibles, la varianza aumentará y el sesgo disminuirá. La tasa relativa de cambio de estas dos cantidades determina si el MSE de los datos de prueba aumenta o disminuye.

Inferencia

- ¿Qué predictores se asocian a la respuesta?

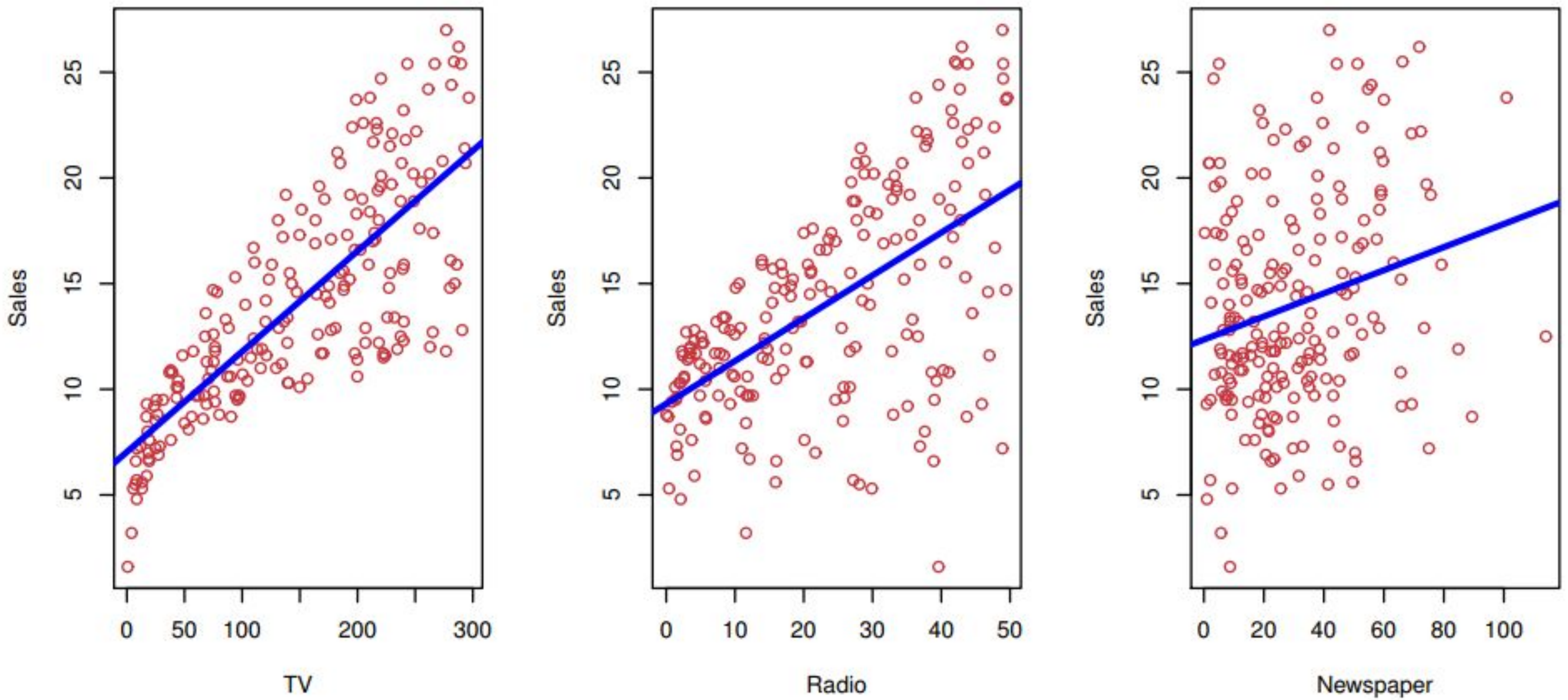
- ¿Cuál es la relación entre la respuesta y cada predictor?





- ¿Puede resumirse adecuadamente la relación entre Y y cada predictor utilizando una ecuación lineal, o es la relación más complicada?

Volviendo al ejemplo de las ventas...



Preguntas posibles...

1. ¿Qué medios de comunicación se asocian a las ventas?

2. ¿Qué medios generan el mayor aumento de ventas?

3. ¿Qué incremento de ventas se asocia a un determinado aumento en la publicidad televisiva?

El objetivo de la Estadística Inferencial es estimar las relaciones entre variables denotadas por magnitudes de efecto, realizar inferencias y predicciones a partir de muestras de una población estadística.

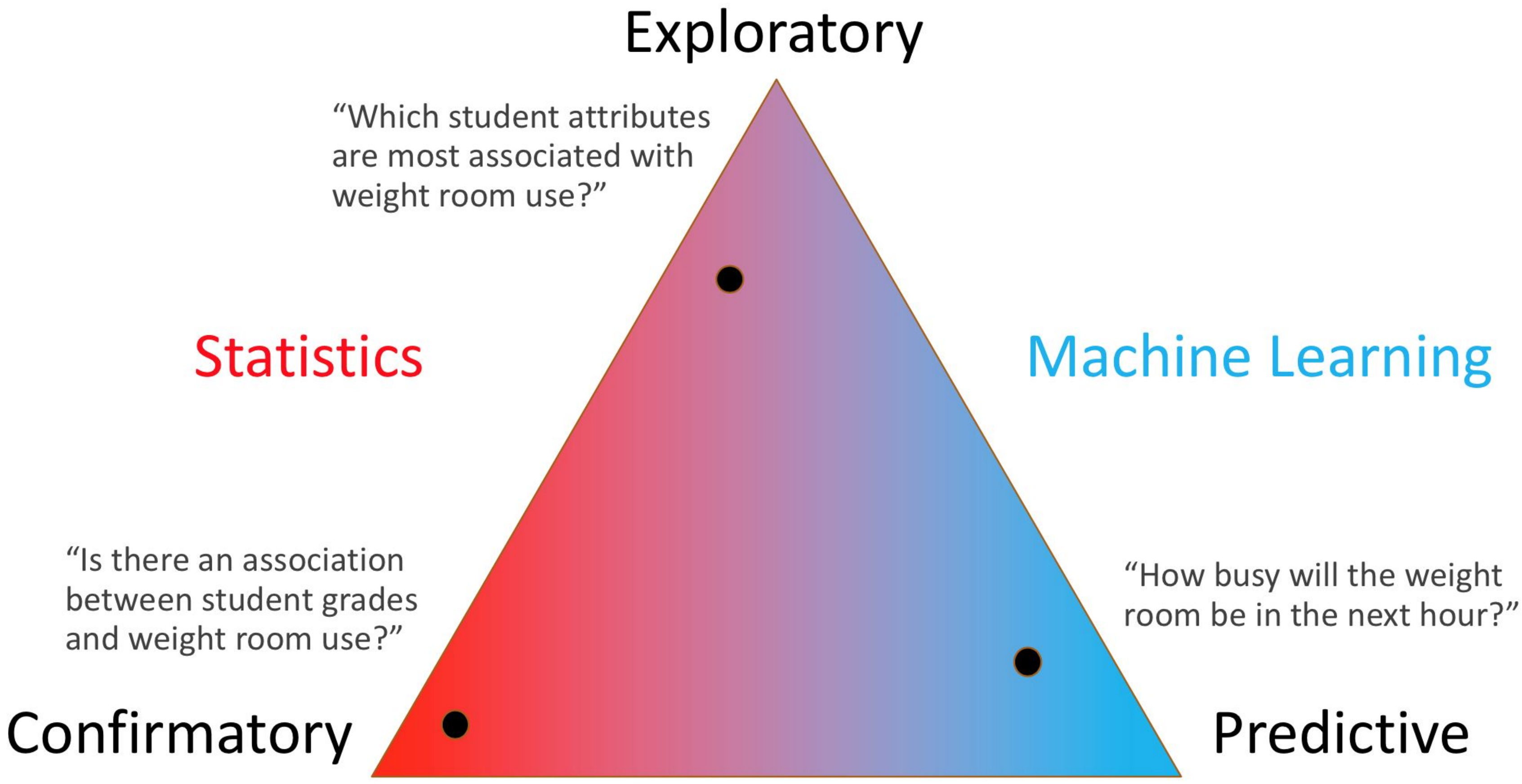
La evidencia (muestra(s)) se obtiene a partir de un diseño experimental/de muestreo que debe ser incorporado en el análisis.

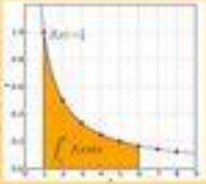
Protocol for conducting and presenting results of regression-type analyses

1. State appropriate questions
2. Visualize the experimental design
3. Conduct data exploration
4. Identify the dependency structure in the data
5. Present the statistical model
6. Fit the model
7. Validate the model
8. Interpret and present the numerical output of the model
9. Create a visual representation of the model
10. Simulate from the model

(Zuur & Ieno 2016)

El objetivo de los análisis debería ser estimar magnitudes de relaciones entre variables (ES), al tiempo que se modelan los principales atributos de los datos (tendencias, variabilidad, etc.) y se validan los modelos ajustados.


Exploratory
“Which student attributes are most associated with weight room use?”
Statistics
Machine Learning
“Is there an association between student grades and weight room use?”
“How busy will the weight room be in the next hour?”
Confirmatory
Predictive



Statistics +  **Machine Learning** = $\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$ **Statistical Learning**

	Statistics	Machine Learning	Statistical Learning
Subfield of...	Mathematics	Computer Science (AI)	Statistics & Machine Learning
Focus on...	Building models with explicitly programmed instructions	Creating systems that learn from data	Sets of tools for modeling and understanding complex data
Purpose	Inferences; Relationships between variables	Optimization; Prediction accuracy	Building statistical models for prediction; understanding data
Prior assumptions about data	Some knowledge about population usually required	None	Some knowledge about population may be required
Dimensionality of data	Usually applied to low-dimensional data	Usually applied to high dimensional data; ML learns from data	Usually applied to high dimensional data
Knowledge overlap	No ML knowledge required	Some stats knowledge usually needed: stats is basis for algorithms	Knowledge of statistics and ML required

Dependiendo de si nuestro objetivo final es la predicción, la inferencia o una combinación de ambas, pueden ser apropiados distintos métodos para estimar *f*. Por ejemplo, los modelos lineales permiten una inferencia relativamente sencilla e interpretable del modelo interlineal, pero pueden no producir predicciones tan precisas como otros métodos.

Modelos Lineales

El Modelo Lineal General: regresión lineal simple como ejemplo inicial (repaso).

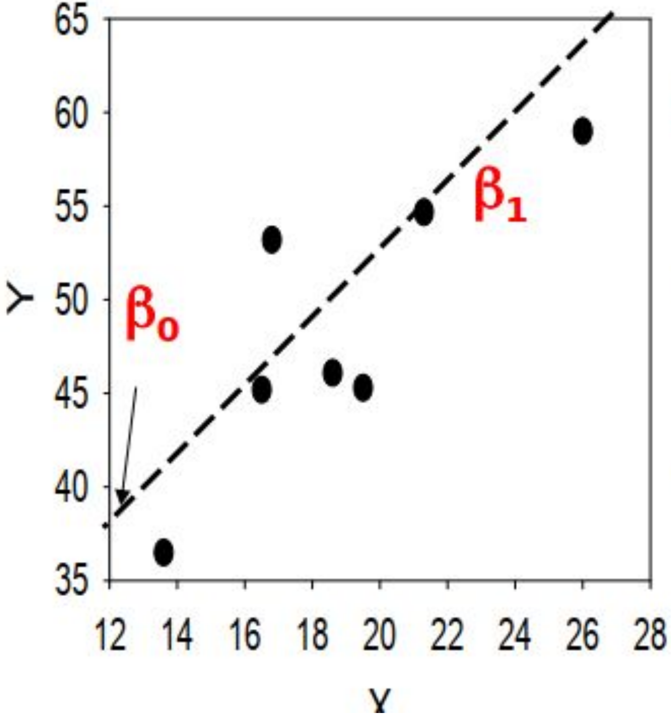
Este modelo probabilístico define:

a) Una relación lineal entre variables Y vs X.

b) Y tiene distribución normal

El modelo de regresión lineal simple:

$$Y_i = \alpha + \beta \times X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2)$$



<u>Los parámetros son</u>:

β_0: intercepto → valor de Y cuando X=0.

β_1: pendiente, tasa de cambio de Y por unidad de cambio de X.

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma_Y^2).$$

Suposiciones del modelo de regresión lineal simple:

1. Y tiene una distribución normal.
2. Las muestras son aleatorias e independientes.
3. X es medida sin error o con un error mínimo.
4. Las varianzas de Y_i para cada X_i son homogéneas.

Práctico:

- Abrimos R y manos a la obra!.
- Abordaremos un ejemplo cuyo Script y datos se encuentran cargados en la página de Git-Hub que les pasamos.

FIN