# Ciencia de Datos con R

# Introducción

Ernesto Mordecki - CMAT, FCIEN, UdelaR

2023

# Los datos

- Tenemos $n$ datos cada uno con $p$ características.
- Simblizamos $x_i \in \mathbb{R}^p$, para $i = 1, \ldots, n$
- En el caso de la base de datos wage (salario) tenemos $n = 3000$ personas en la base de datos con 11 características:

$$\text{edad}, \quad \text{educación}, \quad \text{año}, \quad \ldots$$

# Visualización

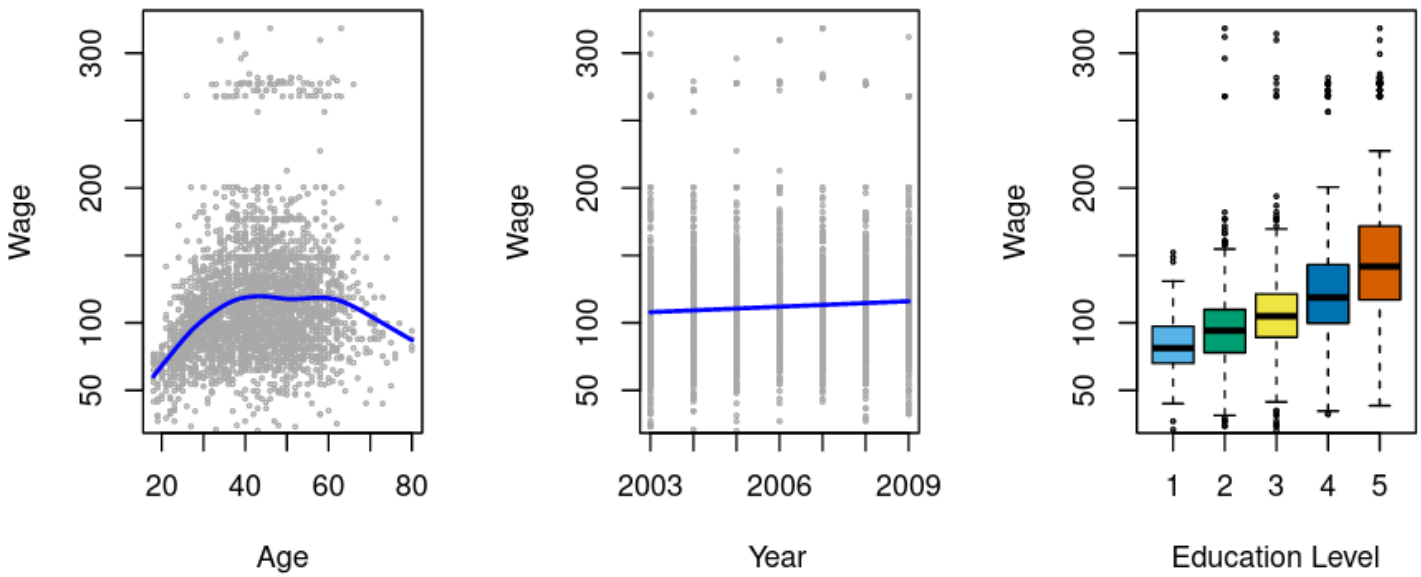


**FIGURE 1.1.** `Wage` *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009. Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

# Notación

- Usamos notación matricial
- Cada fila es un dato con sus $p$ características
- Tenemos $n$ filas

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

- Tenemos un vector

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

que predecir o explicar

- en el ejemplo wage es el salario de cada persona

# Datos y variables aleatorias

- En general mediante letras minúsculas $y_i$ $x_{ij}$ representamos datos
- Mediante letras mayúsculas $Y$, $X_i$ representamos variables aleatoras que tomas los valores de los datos que tenemos
- Si corresponde a realizar un experimento que nos va a dar un dato, mediante $Y$ nos referimos antes de hacer el experimento, no conocemos el dato
- Mediante $y$ nos referimos al dato obtenido
- Es así que atras de todo hay un espacio de probabilidad $(\Omega, \mathcal{B}, \mathbf{P})$.

# Aprendizaje estadístico

- Problema: representar $Y$ como función de $X_1, \ldots, X_p$ predictores
- Buscamos una función $f$ tal que

$$Y = f(X) + \varepsilon.$$

- El $\varepsilon$ significa que aceptamos un cierto error estadístico, o aleatorio.
- Le exigimos dos cosas:
  - Debe ser independiente de los datos
  - Debe ser centrado: $\mathbf{E}(\varepsilon) = 0$

# Predicción VS inferencia

Proponemos un estimador

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

Conocer $Y$ puede tener dos objetivos

- Predicción Tenemos un nuevo dato $x$, queremos predecir $y$
- Esto lo podemos hacer con un mecanismo de caja negra[1]
- Inferencia queremos entender la relación entre los $x$ y los $y$: que características son mas importantes para explicar $Y$, que error se comete, etc.

[1] En la Licenciatura en Matemática tenemos intención de CONSTRUIR estas cajas negras, por lo que deberíamos entender como funcionan

# Calidad del aprendizaje

- Medidimos el error de la estimación mediante

$$\mathbf{E}(Y - \hat{Y})^2$$

- Esta cantidad se llama error cuadrático medio - ECM
- En inglés es el mean square error - MSE
- Tiene dos partes

$$\begin{aligned}\mathbf{E}(Y - \hat{Y})^2 &= \mathbf{E}(f(X) + \varepsilon - \hat{f}(X))^2 \\ &= \mathbf{E}(\hat{f}(X) - f(X))^2 + 2\mathbf{E}((\hat{f}(X) - f(X))\varepsilon) + \mathbf{E}(\varepsilon^2) \quad = \mathbf{E}\end{aligned}$$

- Error reducible, depende del método $\hat{f}$.
- Error irreducible, depende de los datos que usamos.

# Inferencia

- $\hat{f}$ no es una caja negra
- queremos entender el procedimiento:
  - elegir los mejores predictores entre $X_1, \ldots, X_p$
  - Entender relaciones entre predictores y $Y$
  - Entender si puede ser una relación lineal:

$$Y = a_1 X_1 + \cdots + a_p X_p + \varepsilon.$$

  - Este modelo se llama regresión lineal

# Métodos paramétricos VS no paramétricos

- La regresión lineal es un modelo paramétrico
- Tenemos que determinar $a_1, \ldots, a_p$ que viven en $\mathbb{R}^p$
- Podemos pensar relaciones funcionales más elaboradas (polinomios, exponenciales, etc.) que dependen de una cantidad finita de parámetros.
- Esos son llamados métodos paramétrico
- Si me planteo

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

y planteo $f \in C^2(\mathbb{R})$ estoy en un espacio de dimensión infinita, y tengo un método no paramétrico

# Interpretabilidad VS precisión de predicción

Muchas veces buscando presición perdemos interpetabilidad.
Es decir, no sabemos:

- que variables son importantes para explicar $Y$
- cómo se relacionan las variables
- pero se logran predicciones mejores que con los metodos que interpretamos

# Interpretabilidad



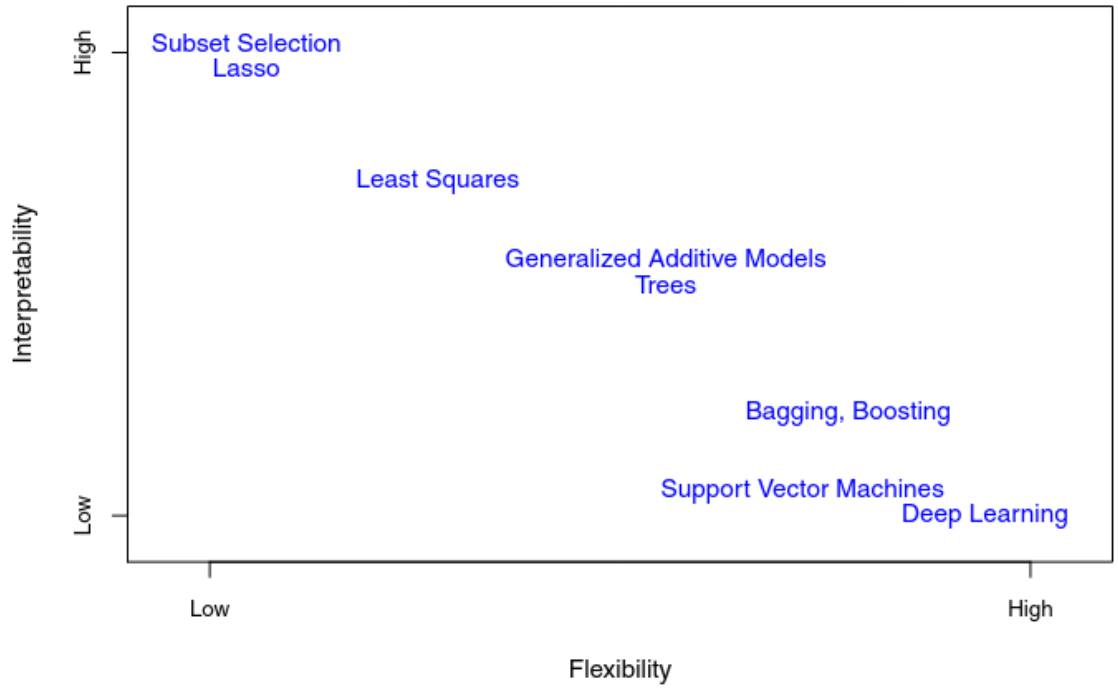


**FIGURE 2.7.** *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

# Aprendizaje supervisado VS no supervisado

- Si es supervisado, conocemos $(y_i, x_i)$ $i = 1, \ldots, n$ e intentamos predecidir o estimar
- Si es no supervisado NO conocemos las $y_i$
- Un problema clásico no supervisado es el agrupamiento o clustering: encontrar grupos de $x$ que tengan $y$ iguales o similares

# Clustering

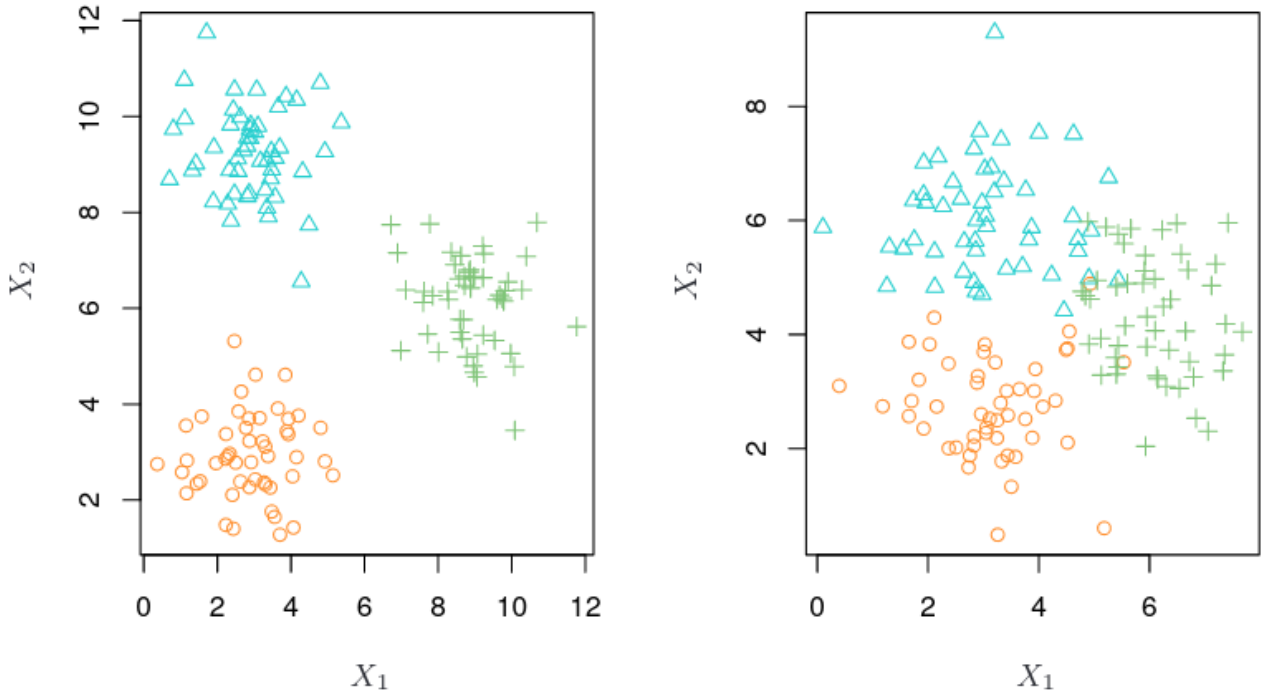


**FIGURE 2.8.** *A clustering data set involving three groups. Each group is shown using a different colored symbol.* Left: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* Right: *There is some overlap among the groups. Now the clustering task is more challenging.*

# Regresión VS clasificación

- Si queremos hallar un valor real numérico de $Y$ tenemos un problema de regresión
- Si queremos hallar un valor discreto entre $\{1, \ldots, K\}$ de $Y$, tenemos clasificación
- Si tenemos es $\{0, 1\}$ tenemos un problema de clasificación binaria.

# Train VS Test

- Una forma de medir la calidad de la estimación $\hat{f}$ es calcular

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2. \tag{1}$$

- Hay un peligro en buscar la mejor $\hat{f}$: los datos que tenemos no representan el todo
- Partimos entonces los datos en dos grupos: datos de entrenamiento - train y datos de testeo - test
- Usamos el conjunto de train para estimar $\hat{f}$
- Medimos el MSE de (1) en ambos conjuntos de datos

# Clustering

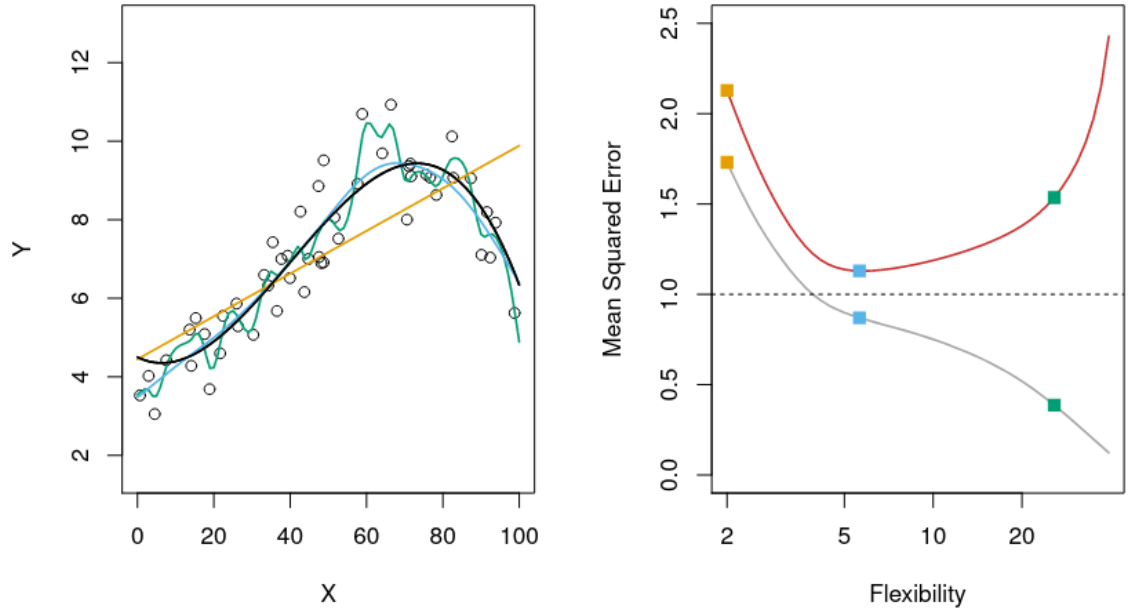


**FIGURE 2.9.** Left: *Data simulated from $f$, shown in black. Three estimates of $f$ are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

En el eje $y$ a la derecha se plotea la cantidad de parámetros que usa el modelo. El primero (naraja) es

$$Y = aX + b + \varepsilon$$

tiene dos parámetros $a$ y $b$.

# Sesgo - Varianza

$$MSE = \mathbf{E}(y - \hat{f}(x))^2 = \mathbf{E}(y^2) - 2\mathbf{E}(y\hat{f}) + \mathbf{E}(\hat{f}^2)$$

donde:

$$\mathbf{E}(y^2) = \mathbf{E}(f + \varepsilon)^2 = f^2 + 2f\mathbf{E}\varepsilon + \mathbf{E}(\varepsilon^2) = f^2 + \mathbf{E}(\varepsilon^2)$$
$$\mathbf{E}\hat{f}^2 = \mathbf{E}(\hat{f} - \mathbf{E}\hat{f} + \mathbf{E}\hat{f})^2 = \mathbf{var}(\hat{f}) + (\mathbf{E}\hat{f})^2$$
$$\mathbf{E}(y\hat{f}) = \mathbf{E}((f + \varepsilon)\hat{f}) = \mathbf{E}(f\hat{f}) + \mathbf{E}(\varepsilon\hat{f}) = \mathbf{E}(f\hat{f}).$$

Entonces

$$MSE = f^2 + \mathbf{E}(\varepsilon^2) + \mathbf{var}(\hat{f}) + (\mathbf{E}\hat{f})^2 - 2\mathbf{E}(f\hat{f})$$
$$= (f - \mathbf{E}\hat{f})^2 + \mathbf{var}(\hat{f}) + \mathbf{E}(\varepsilon^2).$$

- Tenemos el sesgo: por cuanto le erramos
- Tenemos la varianza: cual es la variablidad del método

# Sesgo - Varianza

- El sesgo entonces es el error en el training set
- La varianza es el error en el test set: si es pequeña, al cambiar el conjunto de datos no cambiará mucho