Modelos Estadísticos para la Regresión y la Clasificación
Estadística Descriptiva

Introducción al curso

Mathias Bourel

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

4 de agosto de 2024

Plan

Horarios y Docentes 2024

- Mathias Bourel, responsable y docente de teórico.
- Micaela Long, docente de práctico

- Horarios: Lunes y Miércoles de 8 a 10 (T), Viernes de 8 a 10 (P)

- Salón 101

Sobre la dinámica del curso

- 6 horas semanales (4 de teórico + 2 de práctico) durante 8 semanas
- En el teórico veremos los fundamentos de los diferentes métodos que trataremos en el curso y en el práctico tendrán lista de ejercicios para resolver.
- Si bien cada vez que sea posible incorporaremos el uso de software para ilustrar los distintos conceptos, la idea es que no sea un freno para avanzar con los conceptos que estudiaremos.
- Habrán algunos ejercicios donde será necesario usar un software estadístico (R, Python), solo a los efectos de hacer algunas modelizaciones simples de ilustración de las técnicas vistas.
- Evaluación:
  1. Entrega de ejercicios al final del curso. Fecha y modalidad a definir.
  2. Un parcial final, probablemente durante el segundo periodo de parciales de la Facultad.

Temario y Cronograma semanal tentativo

A grandes rasgos el curso consta de los siguientes tópicos generales de la estadística multivariada: estadística descriptiva y exploratoria, clustering, regresión, y por último clasificación.

- Semana 1 (5 Ago - 9 Ago): Estadística descriptiva, repaso de probabilidad, estimación
- Semana 2 (12 Ago - 16 Ago): Estimación - Regresión Lineal
- Semana 3 (19 Ago - 23 Ago): Regresión Logística
- Semana 4 (26 Ago - 30 Ago): Aprendizaje no supervisado - Clustering
- Semana 5 (2 Set - 6 Set): Aprendizaje no supervisado - Componentes Principales
- Semana 6 (9 Set - 12 Set): Aprendizaje Supervisado - Análisis Discriminante, kNN, CART, SVM
- Semana 7 (16 Set - 19 Set): Aprendizaje supervisado - Combinación de clasificadores y regresores
- Semana 8 (30 Set - 04 Oct): Otros tópicos

Bibliografía

El curso se basa en varios artículos y capítulos de diferentes libros. Los principales libros de referencia son:

- Daniel Peña, *Análisis multivariante de datos*.
- Bernard Flury, *A first course in multivariate statistics*.
- Husson, *Exploratory multivariate analysis by example using R*.
- James, Witten, Hastie y Tibshirani *An introduction to statistical learning with applications in R*.
- James, Witten, Hastie y Tibshirani *An introduction to statistical learning with applications in Python*.
- Andrew Wolf, *Machine Learning Simplified: A Gentle Introduction to Supervised Learning*
- Hadley Wickham y Garrett Grolemund, *R para Ciencia de Datos*

Plan

Estadística descriptiva y exploratoria

Análisis exploratorio de datos

Es un enfoque para analizar conjuntos de datos, para resumir sus características principales, casi siempre mediante gráficos que ayudan a la visualización.

Con o sin modelos estadísticos

Se puede usar o no un modelo estadístico, pero sobre todo es para ver qué nos pueden decir los datos más allá del modelo formal que utilicemos después.

Ayuda a:

- entender la estructura de los datos, detectar errores en ellos.
- formular hipótesis que podrían conducir a la recopilación de nuevos datos y experimentos. Pero no responde a problemas causales y/o de inferencia, sino que da pistas.
- verificar los supuestos requeridos para el ajuste de modelos, manejar datos faltantes y hacer transformaciones de variables según sea el caso.

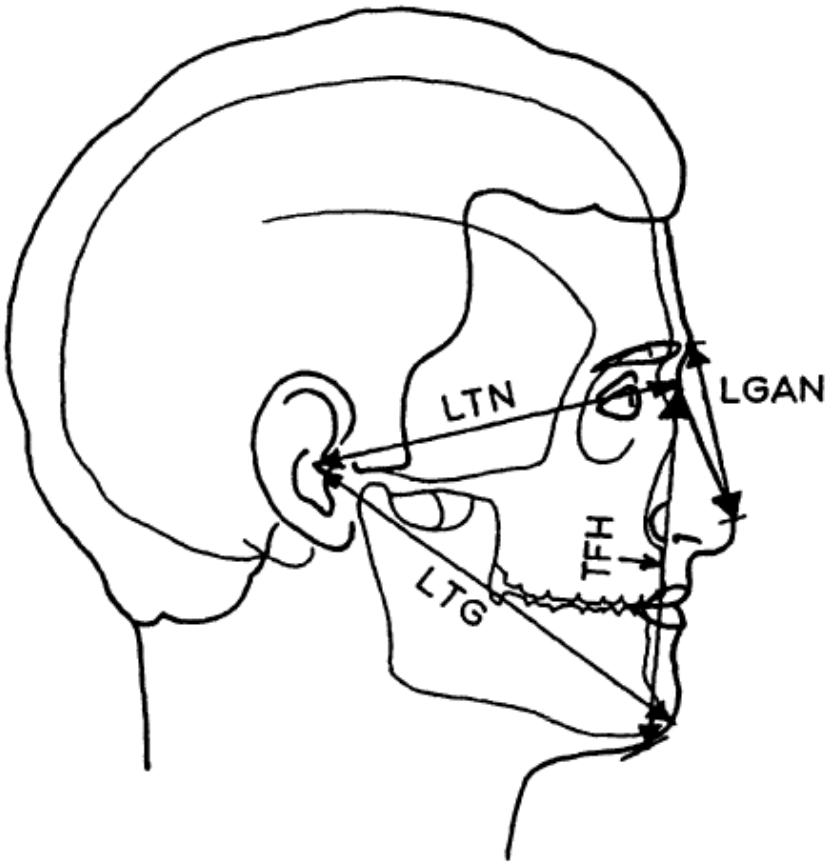
Ejemplo: Máscaras de protección

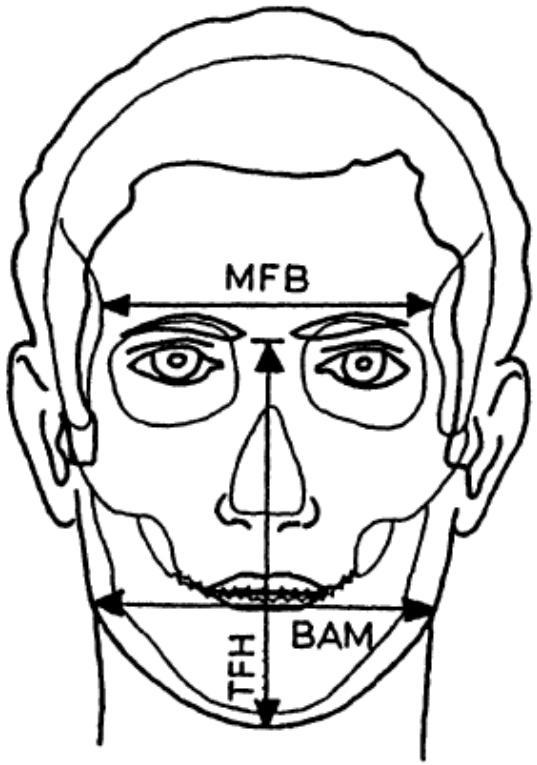
Descripción del conjunto de datos

A mediados de la década de 1980, el gobierno suizo decidió que se necesitaban nuevas máscaras de protección para los miembros de su ejército.
En total 200 miembros del ejército fueron muestreados, todos de 20 años de edad, y se obtuvieron medidas de las 6 variables que los fabricantes de las máscaras consideraron más importantes.

MFB =	anchura frontal mínima
BAM =	amplitud de angulus mandibulae
TFH =	altura facial verdadera
LGAN =	longitud desde glabela hasta el ápice nasal
LTN =	longitud de tragión a nasión
LTG =	longitud de tragión a gnathion

Ejemplo: Máscaras de protección





Ejemplo: Máscaras de protección

MFB	BAM	TFH	LGAN	LTN	LTG
113.2	111.7	119.6	53.9	127.4	143.6
117.6	117.3	121.2	47.7	124.7	143.9
112.3	124.7	131.6	56.7	123.4	149.3
116.2	110.5	114.2	57.9	121.6	140.9
112.9	111.3	114.3	51.5	119.9	133.5
104.2	114.3	116.5	49.9	122.9	136.7
110.7	116.9	128.5	56.8	118.1	134.7
105	119.2	121.1	52.2	117.3	131.4
⋮	⋮	⋮	⋮	⋮	⋮

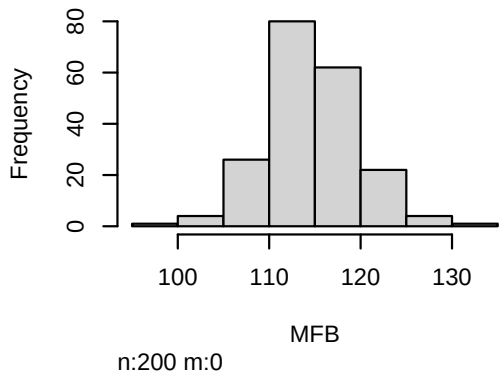
La tabla muestra los datos, que razonablemente se pueden considerar como una muestra aleatoria de la "población" de todos los varones suizos sanos de 20 años. Por tanto, estudiaremos la distribución conjunta de $p = 6$ variables, basándonos en $N = 200$ observaciones.

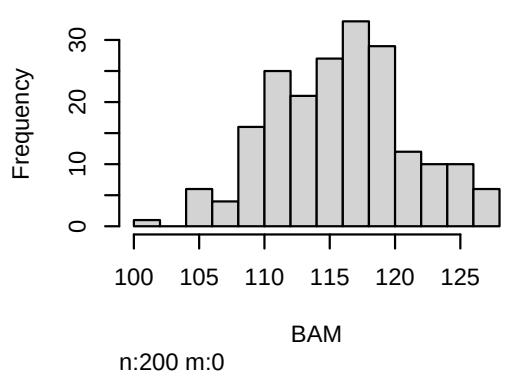
Ejemplo: Máscaras de protección

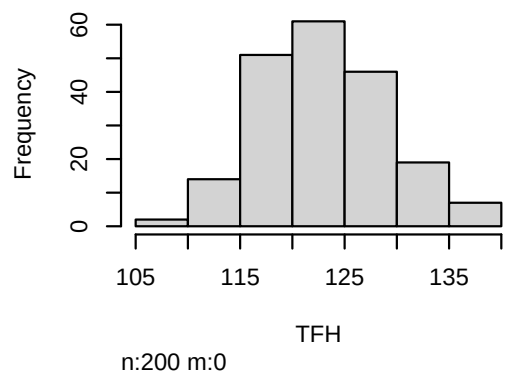
Se suele usar resúmenes numéricos, como la media, la mediana, los cuartiles y los extremos:

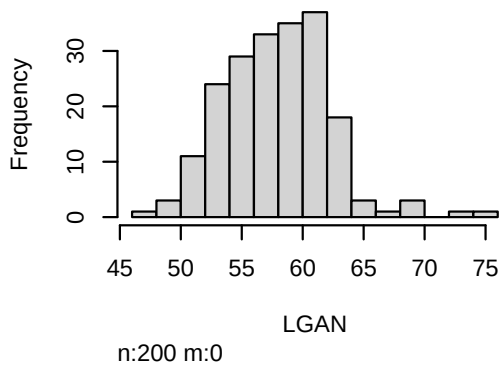
	MFB	BAM	TFH	LGAN	LTN	LTG
Min.	96.8	100.6	108.7	47.70	112.1	122.6
1st Qu.	111.2	111.9	118.0	54.90	119.9	135.1
Median	114.5	116.2	122.7	57.95	122.2	139.0
Mean	114.7	115.9	123.1	57.99	122.2	138.8
3rd Qu.	118.3	119.1	127.7	60.80	124.8	142.8
Max.	130.5	127.8	139.1	74.20	134.7	153.3

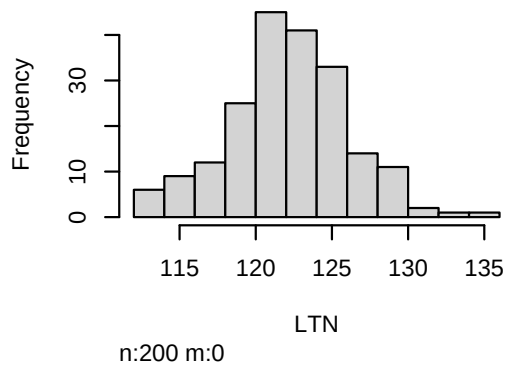
Ejemplo: Máscaras de protección

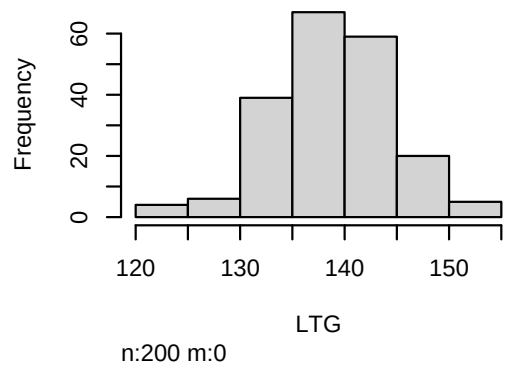




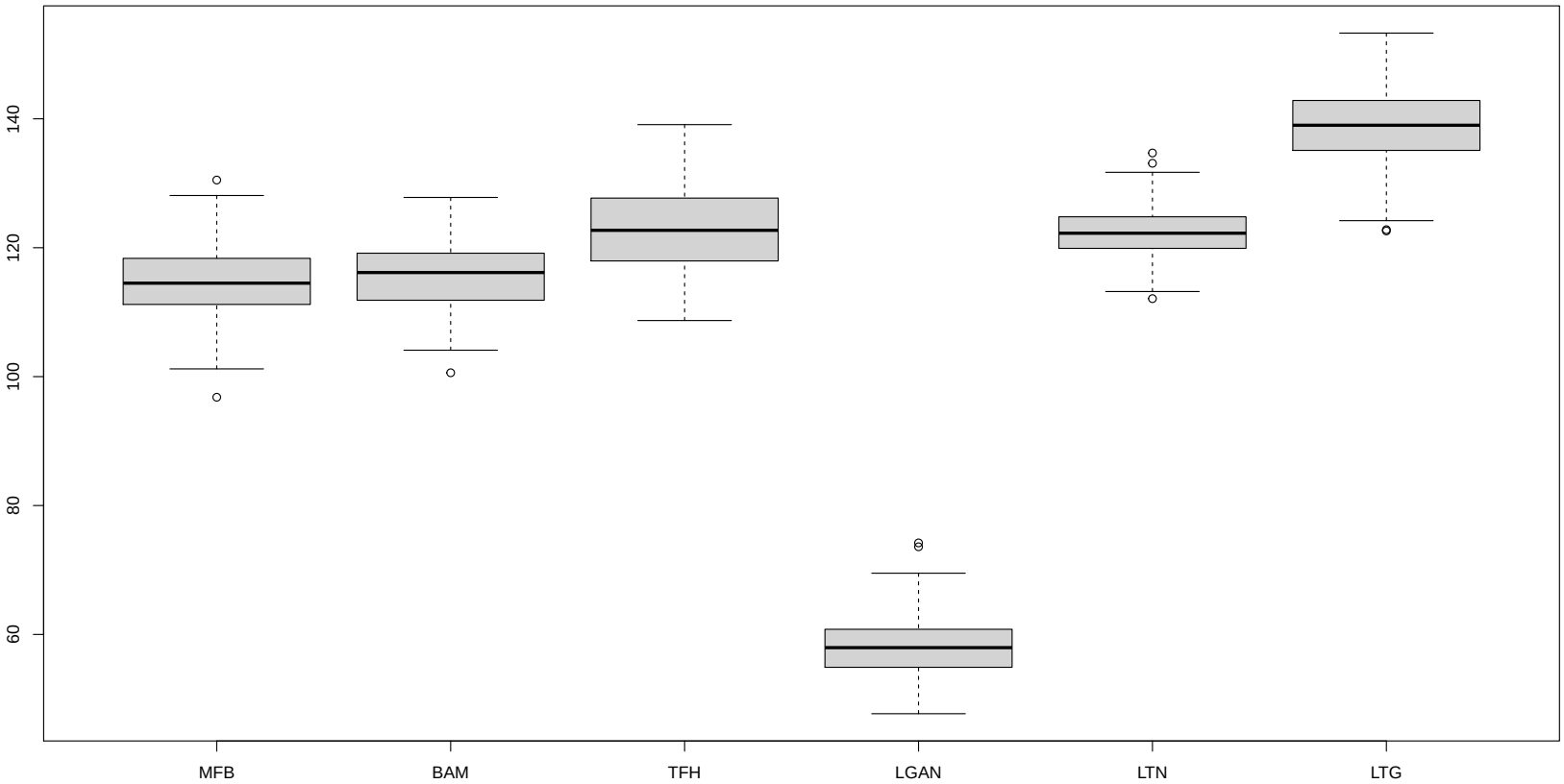




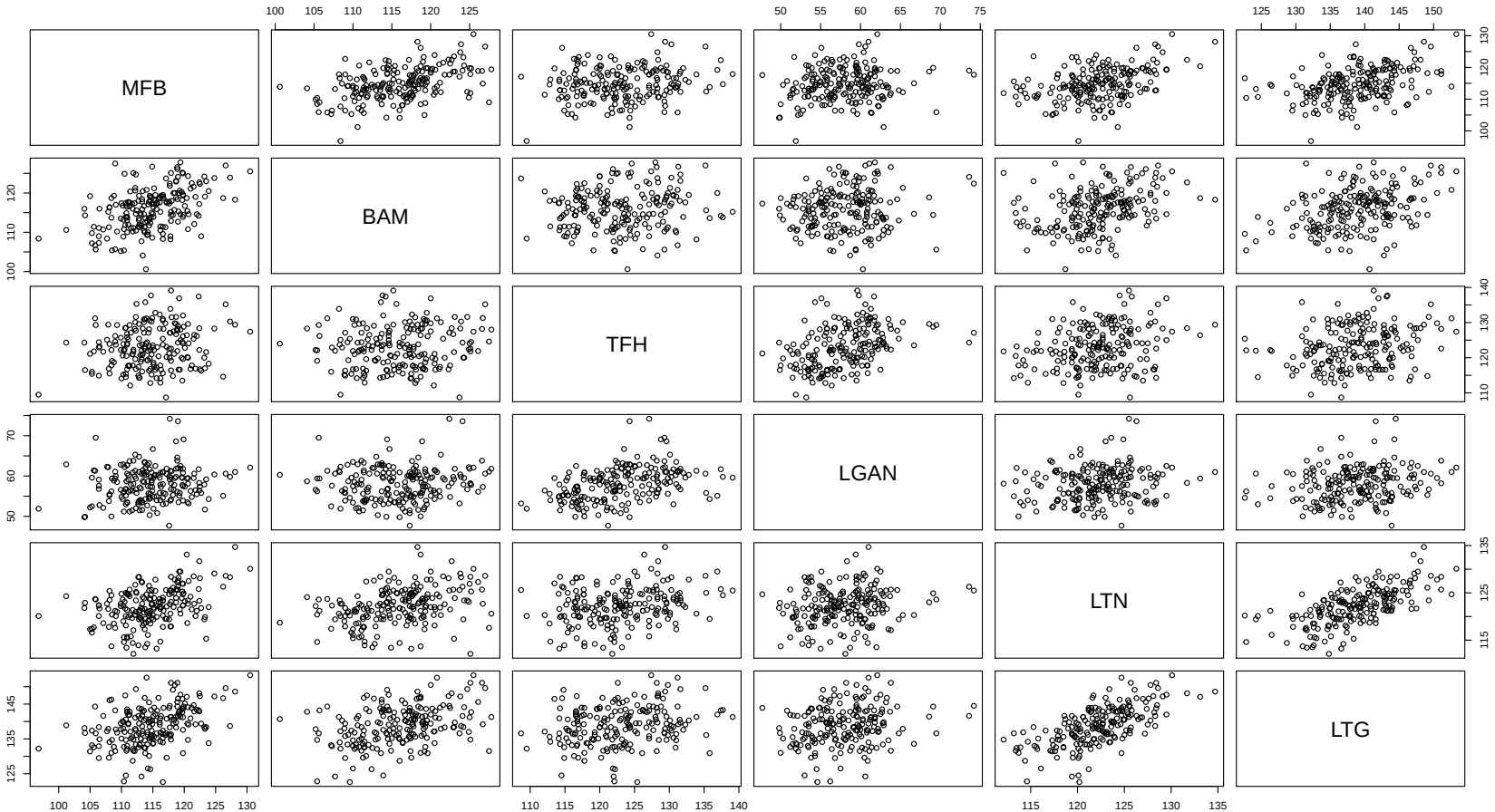




Ejemplo: Máscaras de protección



Ejemplo: Máscaras de protección



Ejemplo: Máscaras de protección

En el estudio de la dependencia dos a dos entre las variables se suele calcular la matriz de correlaciones:

	MFB	BAM	TFH	LGAN	LTN	LTG
MFB	1.000	0.466	0.175	0.134	0.402	0.414
BAM	0.466	1.000	0.093	0.093	0.348	0.388
TFH	0.175	0.093	1.000	0.414	0.259	0.238
LGAN	0.134	0.093	0.414	1.000	0.176	0.210
LTN	0.402	0.348	0.259	0.176	1.000	0.656
LTG	0.414	0.388	0.238	0.210	0.656	1.000

Ejemplo: Máscaras de protección

Objetivo

Usar la información contenida en este conjunto de datos para diseñar las máscaras. Por ejemplo, un conjunto de tres o cuatro tipos de máscaras que representen mejor a la población.

Análisis de componentes principales: reducción de la dimensión

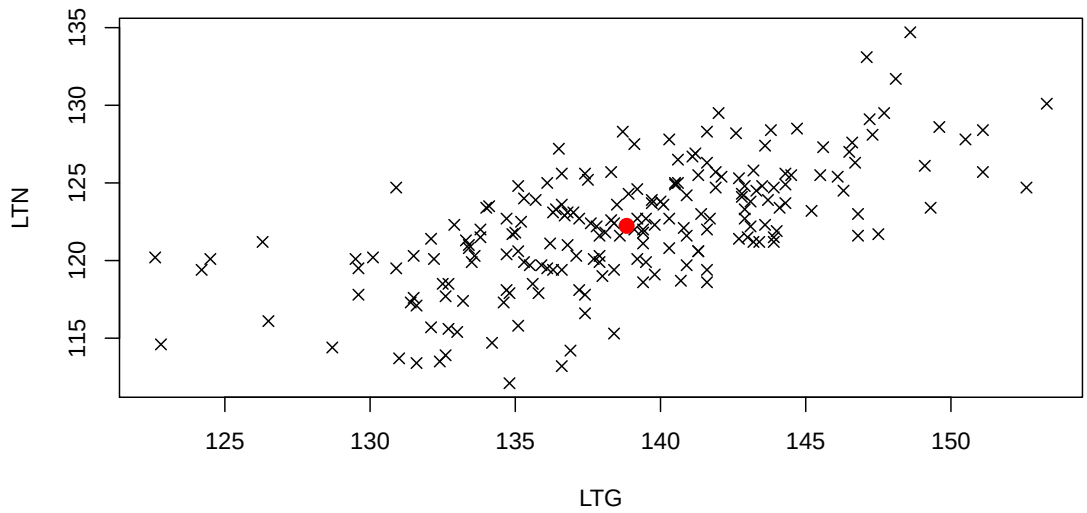
Sería fantástico si pudiéramos construir y comprender diagramas de dispersión de seis dimensiones para detectar la estructura en el conjunto de datos...
El análisis de componentes principales es una técnica de reducción de la dimensión que intenta satisfacer los siguientes criterios:

- Distorsionar lo menos posible la *forma* original de la nube de puntos en el espacio 6-dimensional.
- Cambiar a un nuevo sistema de coordenadas (serían como nuevas variables) que sean lo menos correlacionadas posible.

Este análisis también nos provee de una medida de la calidad de la representación obtenida.

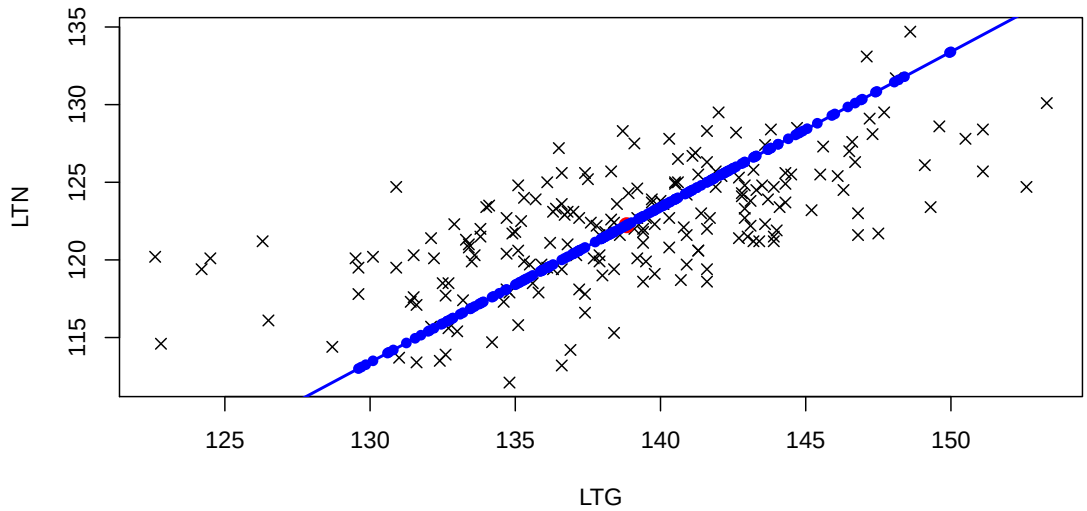
Ejemplo: Máscaras de protección

Para simplificar, consideremos solamente las variables LTG y LTN. Si fueramos a fabricar una sola máscara de gas, lo razonable sería considerar el promedio (el punto rojo en la gráfica).



Ejemplo: Máscaras de protección

Algo mejor sería reducir la dimensión a 1, proyectando los puntos sobre una recta (como en la recta azul del dibujo). Buscamos así la recta que distorsione los menos posible las distancias entre los puntos.

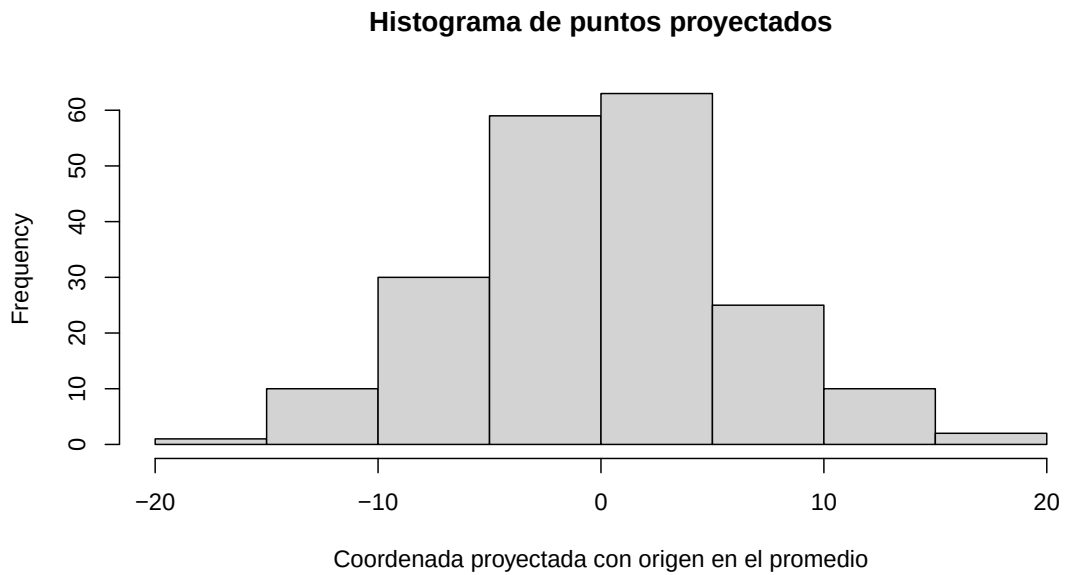


Ejemplo: Máscaras de protección

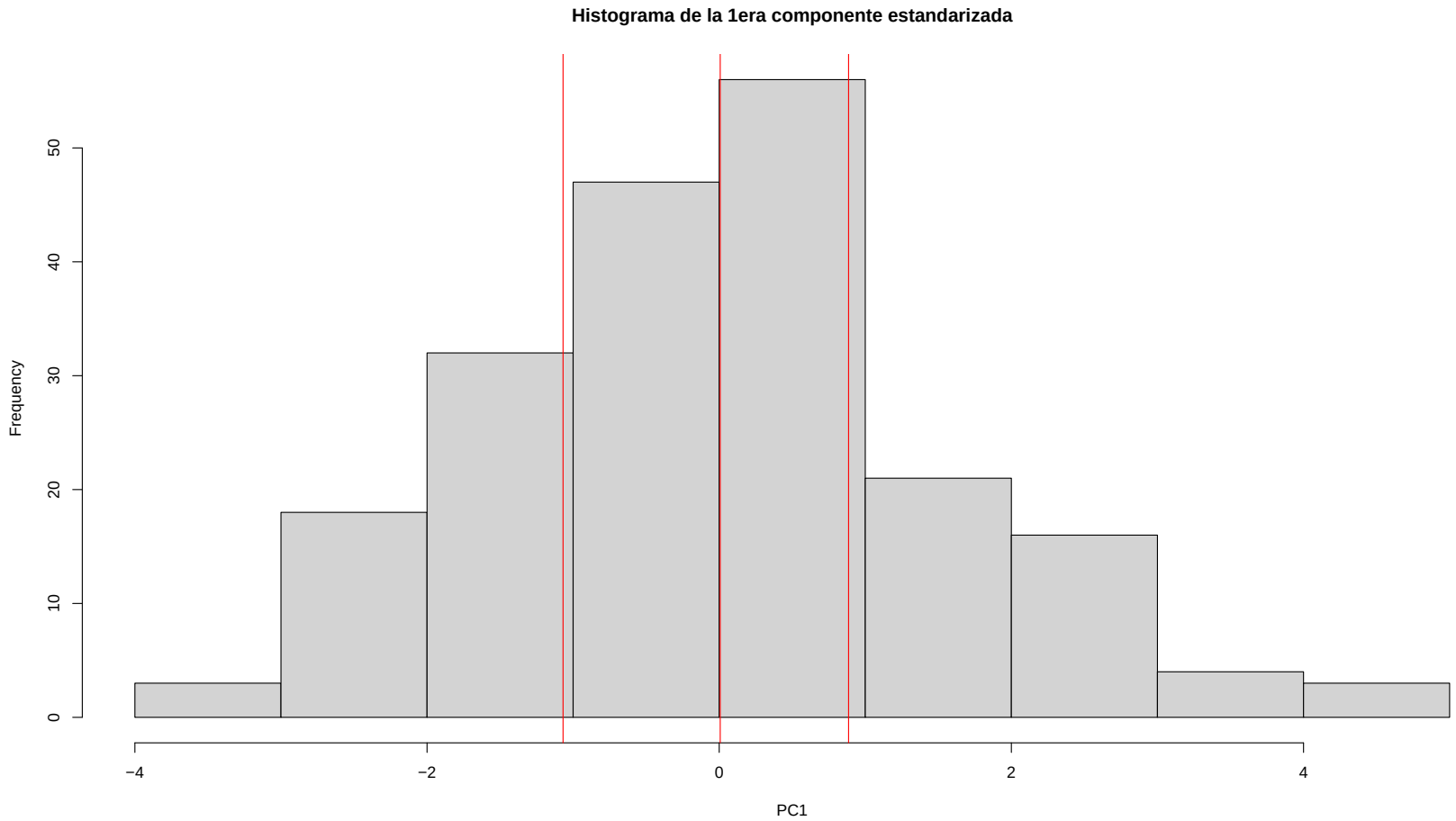
Así buscamos la recta r que realice el siguiente mínimo:

$$\min_r \left\{ \sum_{x,y} |d(x,y) - d_r(x,y)| \right\}$$

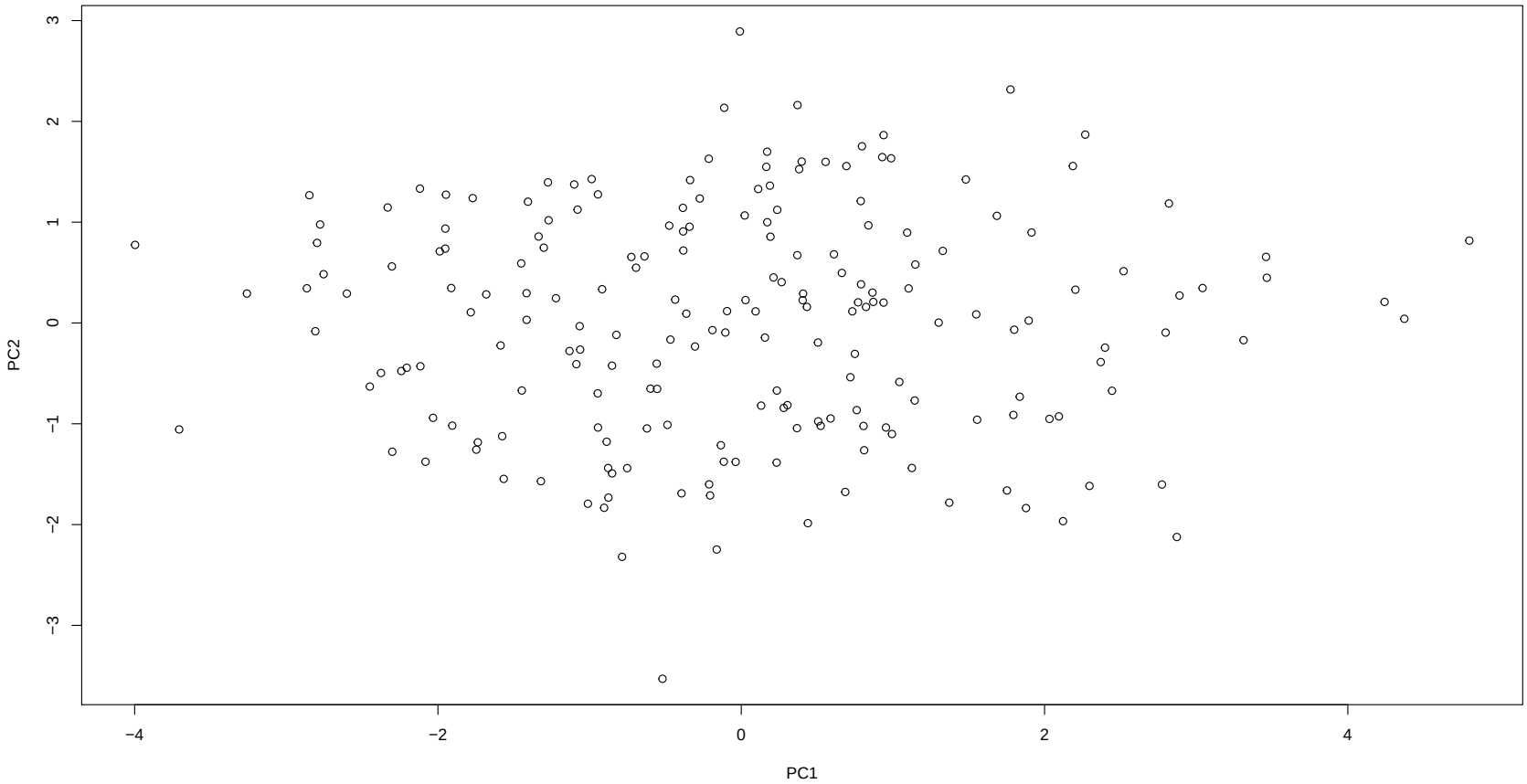
En esta recta, con origen en el promedio, los datos se distribuyen de la siguiente manera:



Ejemplo: Máscaras de protección



Ejemplo: Máscaras de protección



Un problema de agrupamiento

Ejemplo: Longitud de las alas de los pájaros

Descripción del conjunto de datos

La captura y marcado de aves es un método importante para estudiar su migración. Al capturar un animal, al biólogo le gustaría obtener ciertas medidas, sin molestar demasiado al ave.
En un experimento, se capturaron y marcaron bisbitas de agua (Anthus spinoletta). Los datos brutos obtenidos de N = 381 aves se resumen en la siguiente tabla. Hay una única variable: la longitud del ala, en milímetros.

Ejemplo: Longitud de las alas de los pájaros

Wing.Length	Frequency
82	5
83	3
84	12
85	36
86	55
87	45
88	21
89	13
90	15
91	34
92	59
93	48
94	16
95	12
96	6
97	0
98	1

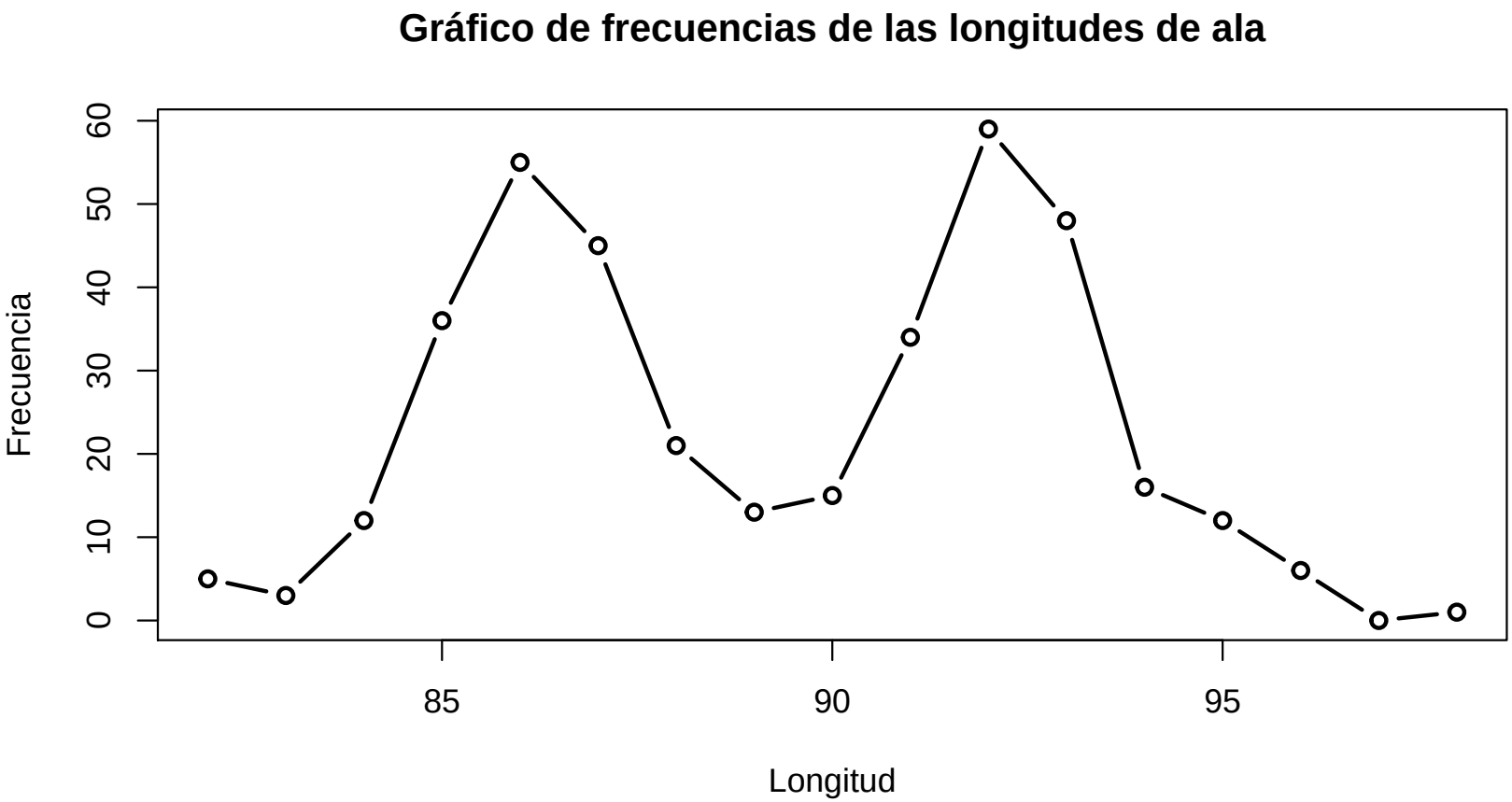


Gráfico de frecuencias de las longitudes de ala
Frecuencia
0
10
20
30
40
50
60
85
90
95
Longitud

Ejemplo: Longitud de las alas de los pájaros

Interpretación del gráfico de frecuencias

Una característica sorprendente de los datos es que hay muchas aves con una longitud de ala en el rango de 85 a 87 mm y nuevamente muchas en el rango de 91 a 93 mm, con un valle "distinto"de baja frecuencia en el medio.
Es probable que la mayoría de las aves más grandes sean machos y la mayoría de las aves más pequeñas sean hembras, pero no se identificó el género de las aves en el momento de la investigación para evitar dañarlas.

Objetivo: Clustering

Nos gustaría establecer una regla que nos permita asignar cada espécimen a uno de los dos grupos, manteniendo la probabilidad de clasificación errónea lo más pequeña posible. Sería útil dividir el conjunto de datos en subconjuntos etiquetados como "probablemente macho" y "probable hembra", respectivamente.

Clustering o Análisis de conglomerados

Los métodos para dividir conjuntos de datos en subgrupos homogéneos se denominan comúnmente análisis de conglomerados.
En la terminología de machine learning este tipo de problemas se denomina *aprendizaje no supervisado*.

En el ejemplo de las aves

Un análisis de conglomerados normalmente encontraría un punto de corte o valor crítico c para la longitud del ala, declararía todas las aves con una longitud de ala menor a c como un grupo y aquellas con una longitud de ala mayor a c como el segundo grupo.

Un modelo probabilístico para clustering

Una forma conveniente de modelar esta situación matemáticamente es pensar en la pertenencia a un grupo como una variable aleatoria.

En el ejemplo de las aves

Esto significa que realmente estamos analizando dos variables, a saber, $X =$ género e $Y =$ longitud del ala. Solo se nos dan observaciones de Y, pero no es posible comprender la distribución de Y sin algún conocimiento de cómo Y depende de X.

Un problema de predicción (regresión)

Ejemplo: La amargura del jugo de naranja

Descripción del conjunto de datos

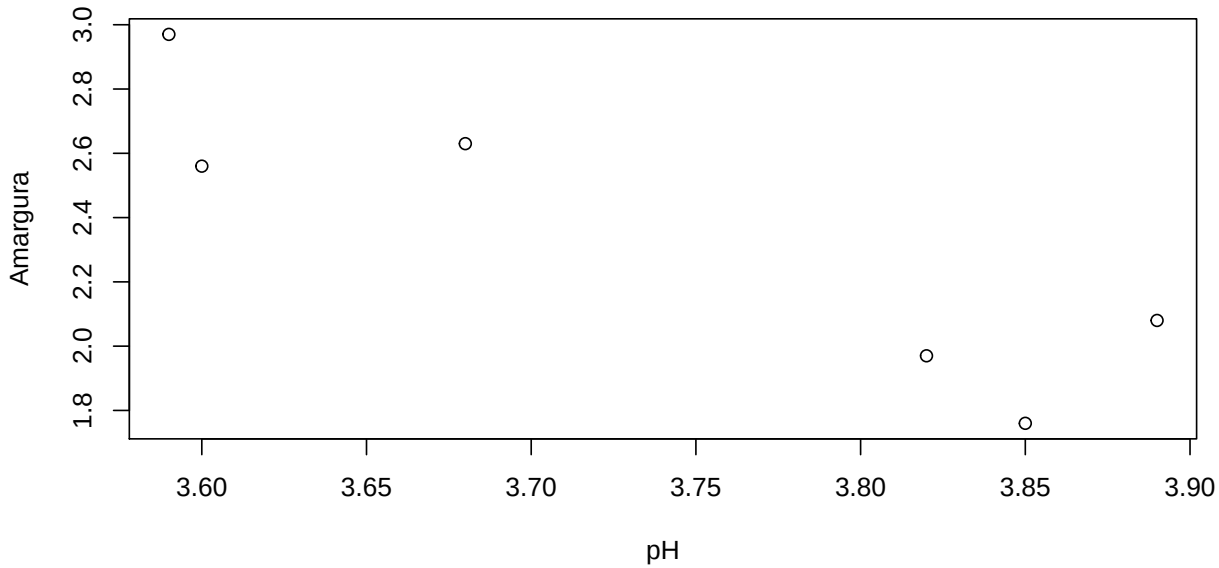
Un estudio pretende predecir la amargura al gusto de diferentes tipos de jugo de naranja, a partir de medidas tomadas en el laboratorio del pH de cada muestra.
El conjunto de datos consta entonces de dos variables, $X =$ pH e $Y =$ a la amargura, una medida subjetiva tomada por una persona en una escala numérica determinada.
Vamos a estudiar entonces la distribución conjunta de $p = 2$ variables, X e Y, a partir de $N = 6$ observaciones.

Objetivo

El objetivo es comprender la relación entre una medida objetiva X y otra subjetiva Y. El foco está en lograr *predecir* la *variable de respuesta* Y a partir de la *variable explicativa* X. Notar la asimetría en el rol que las variables juegan en este ejemplo.

Ejemplo: La amargura del jugo de naranja

pH	Bitterness
3.59	2.97
3.89	2.08
3.85	1.76
3.60	2.56
3.82	1.97
3.68	2.63



Ejemplo: La amargura del jugo de naranja

Regresión lineal

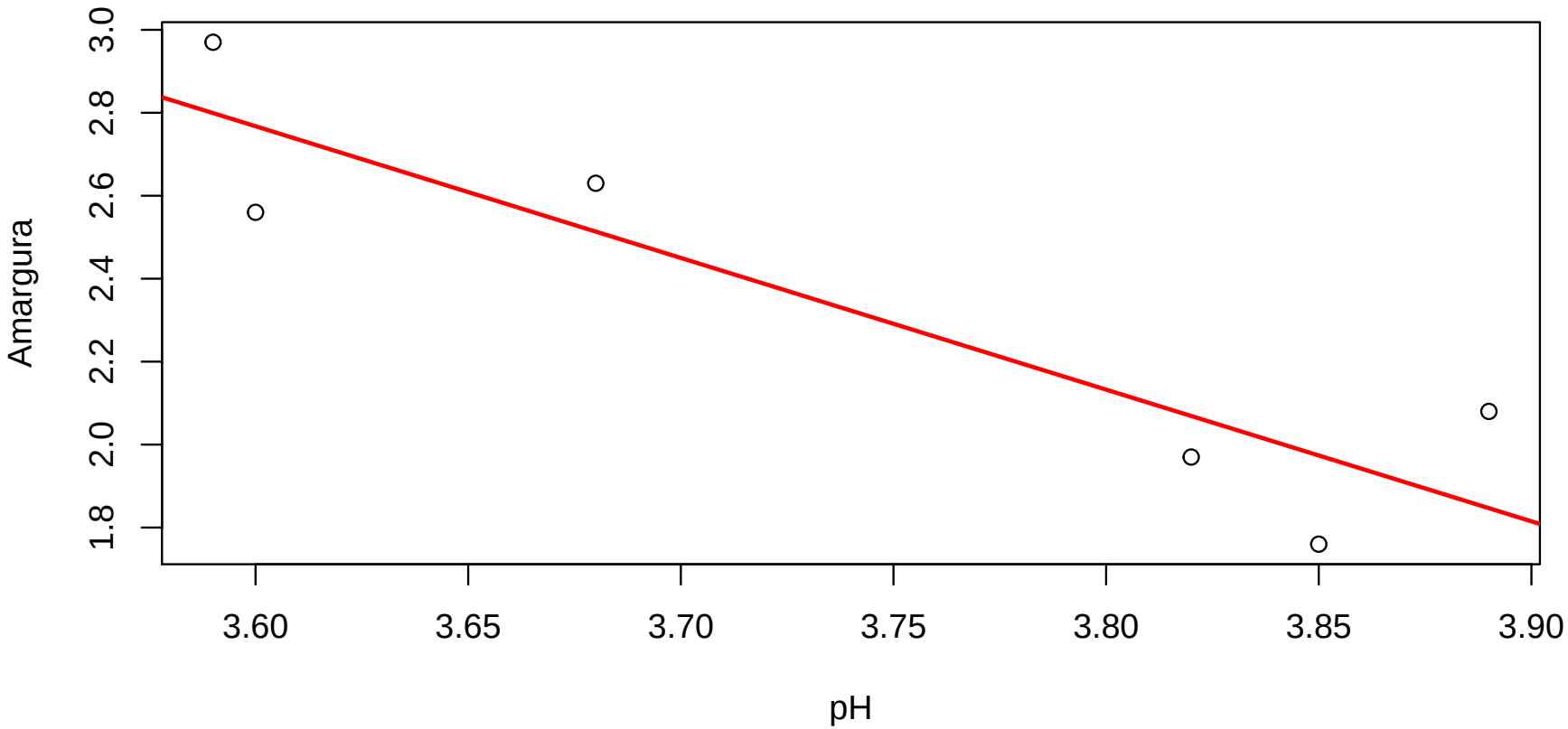
Una forma de modelar matemáticamente este problema es mediante el *análisis de regresión*. La función de regresión es por definición:

$$R(x) = \mathbf{E}(Y|X = x)$$

y representa una predicción razonable de Y para el valor $X = x$. Cuando esta función es *lineal*, la regresión se dice lineal.

Correlación y mínimos cuadrados

La regresión está relacionada con la correlación (de hecho la antecede históricamente). Se puede interpretar como la proyección ortogonal de Y sobre la recta afín generada por X ($bX + a$) y es por este motivo que se generalmente su estimación se basa en el método de mínimos cuadrados.


Amargura
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.60
3.65
3.70
3.75
3.80
3.85
3.90
pH

Otro problema de predicción (clasificación)

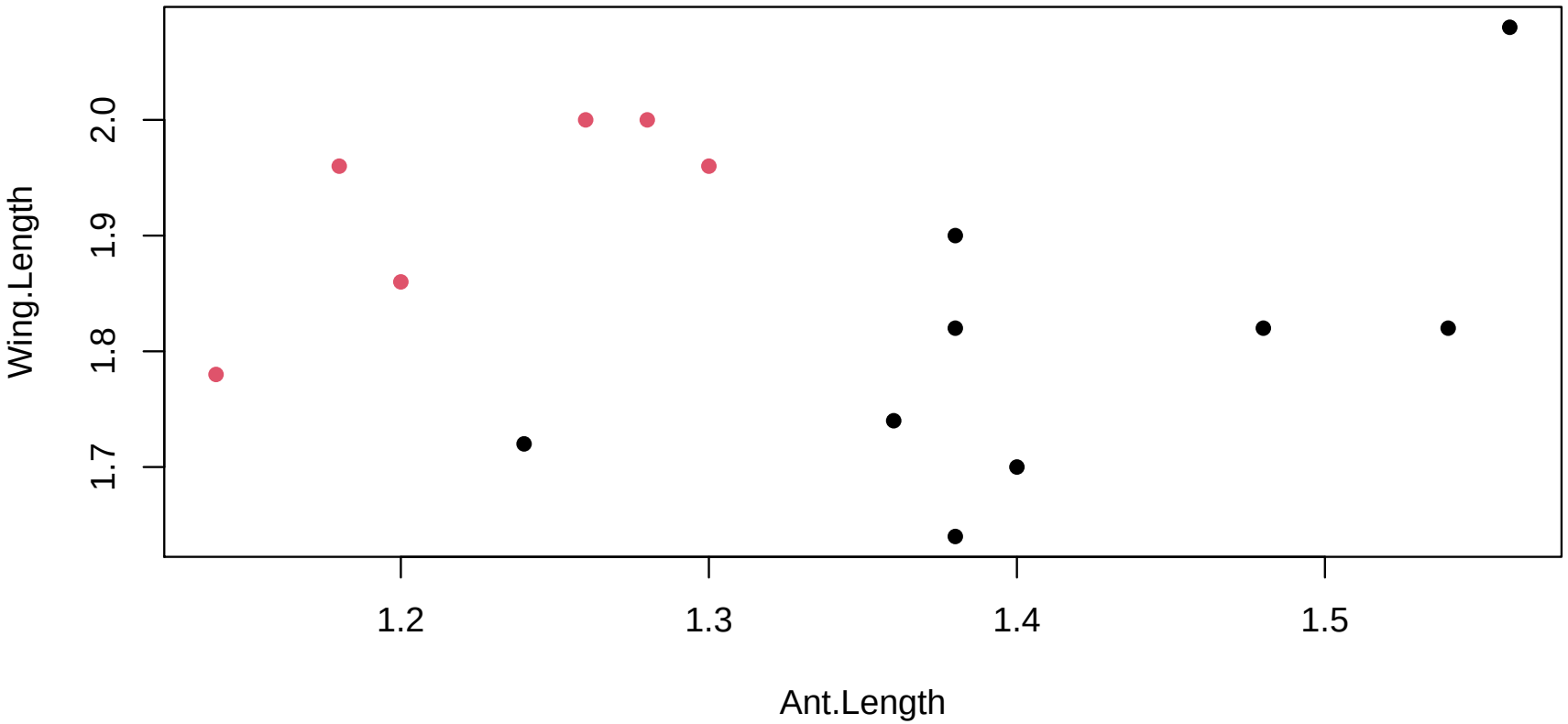
Ejemplo: Clasificación de mosquitos (midges)

Descripción del conjunto de datos

Un biólogo describió dos especies recién descubiertas de mosquitos depredadores, midges en inglés, Amerohelea fasciata (Af) y Amerohelea pseudo fasciata (Apf). Los midges son pequeños insectos parecidos a mosquitos.

Objetivo: clasificación

Debido a que las dos especies son similares en apariencia, es útil para el biólogo poder clasificar un espécimen como Af o Apf basándose en características externas que sean fáciles de medir. La intención es clasificarlas en base a sus medidas de longitud de antena y alas. La pregunta es entonces, ¿podemos idear una regla automática que clasifique, con el menor error posible, un *nuevo* espécimen de midge basándonos en sus medidas de antena y alas?


Wing.Length
1.7
1.8
1.9
2.0
1.2
1.3
1.4
1.5
Ant.Length

Ejemplo: Clasificación de mosquitos (midges)

Clasificación en estadística

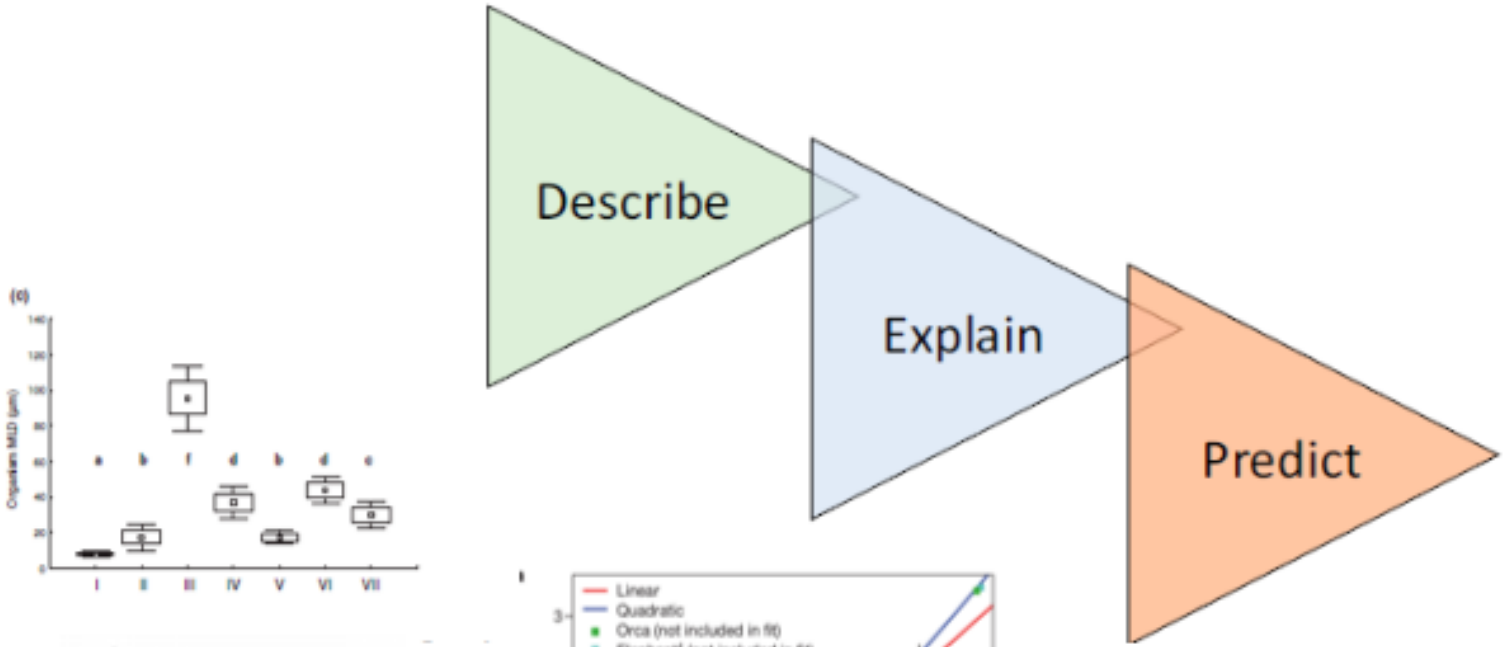
El problema de clasificación es parecido al de clustering, en el sentido que queremos separar grupos, salvo que en este caso disponemos de un conjunto de datos que sirve de entrenamiento y sabemos a que grupo pertenece cada individuo. El objetivo no es agrupar los datos disponibles (pues ya conocemos a qué grupo pertenecen), sino determinar la clase de un futuro individuo que no forma parte de los datos disponibles. En la terminología de machine learning este tipo de problemas se denomina de *aprendizaje supervisado*.

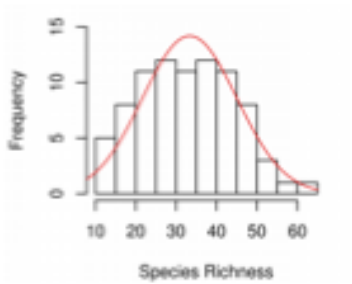
Diferentes métodos de clasificación

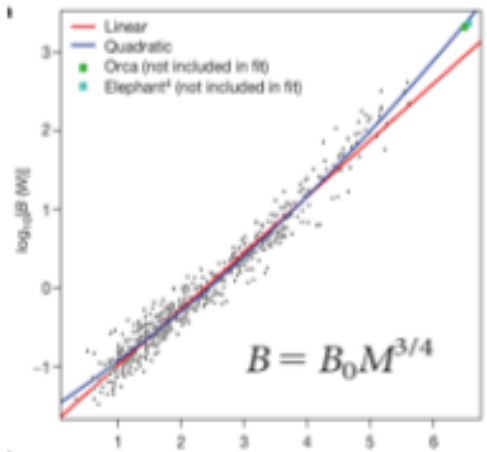
En el curso veremos algunos métodos de clasificación, por ejemplo:

- Regresión logística
- Arboles de clasificación
- Support Vector Machines
- k-vecinos más cercanos
- Naive Bayes

Hoja de ruta del análisis estadístico







Kolokotrones et al 2010

Leo Breiman, 2001, The Two Cultures

Leo Breiman. Statistical Learning: The Two Cultures. Statistical Science. 16 (3): 199-231, 2001.

Leo Breiman, 2001, The Two Cultures

- Hay dos culturas en el uso de modelos estadísticos para llegar a conclusiones a partir de datos.
- Una asume que los datos son generados por un determinado modelo estocástico de datos (*La cultura del modelado de datos*).
- La otra utiliza modelos algorítmicos y trata el mecanismo de los datos como desconocido (cultura del modelado algorítmico).
- La comunidad estadística ha estado comprometida con el uso casi exclusivo de modelos de datos. Este compromiso ha llevado a una teoría irrelevante, conclusiones cuestionables, y ha impedido a los estadísticos trabajar en una gran variedad de interesantes problemas actuales.

Modelización estadística

- Las técnicas de Aprendizaje Estadístico pueden ayudar a resolver los problemas que surgen con frecuencia al modelizar un problema ecológico, un fenómeno económico, una situación médica, una situación climática, etc.
- Idea: a partir de un conjunto de datos (de entrenamiento), construir y entrenar un modelo matemático f que permita, dada una nueva observación, predecir la categoría a la que pertenece o algún valor de salida relevante. El predictor f se construye generalmente sin hacer ningún supuesto en cuanto a la distribución o la naturaleza del conjunto de datos.
- Si Y es la variable de respuesta:

$$\text{Modelo } Y = f(X) + \epsilon$$

$$\text{predicción: } \hat{Y} = \hat{f}(X)$$

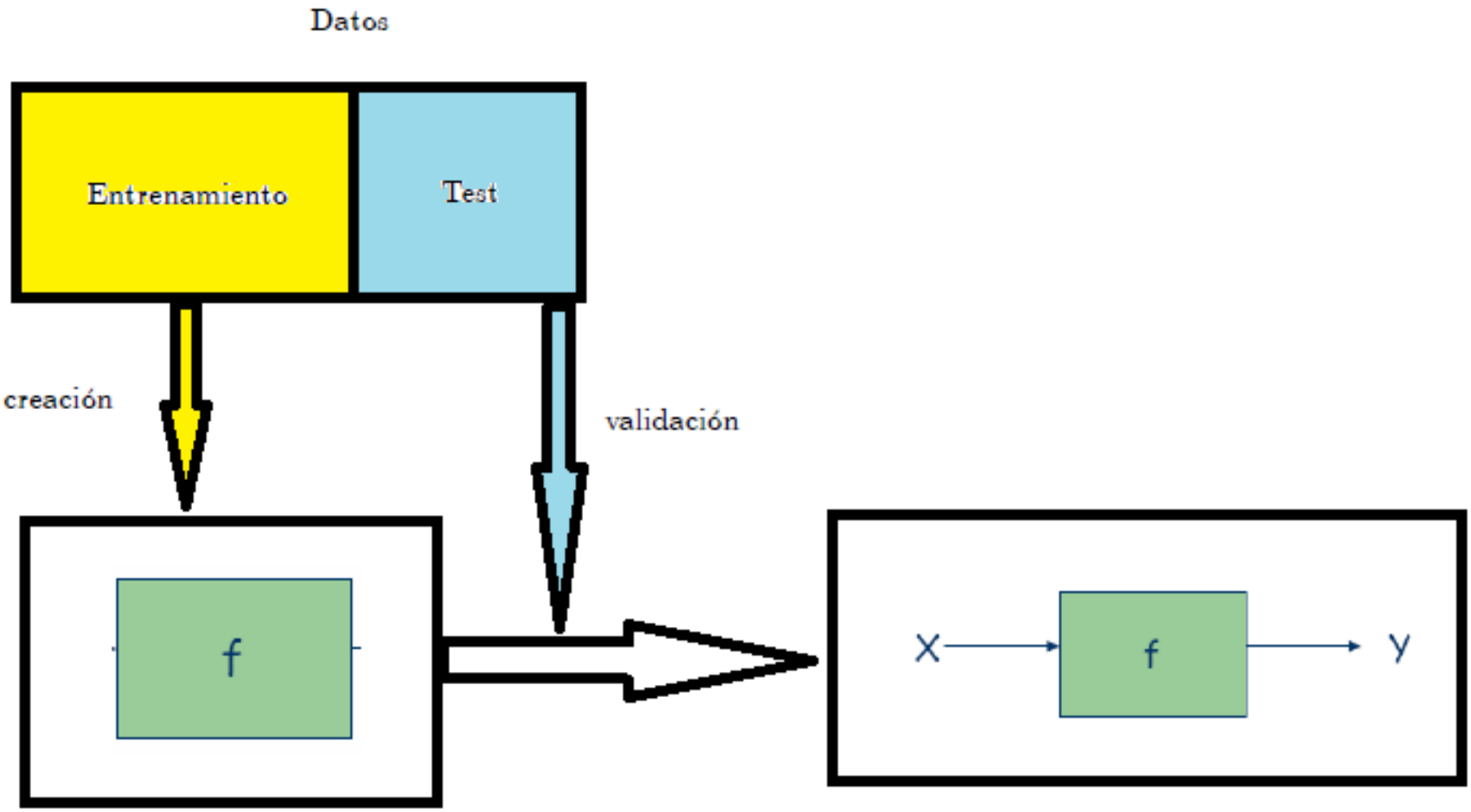
$$\text{vamos a querer: } \hat{Y} \approx Y$$

  - Cultura de modelado de datos: f tiene una forma determinada (regresión lineal o logística) y estimamos los parámetros a partir de los datos. Se trabaja para el modelo. La validación se refiere (generalmente) a la bondad del ajuste del modelo a los datos.
  - Cultura de modelización algorítmica: f es un algoritmo. La validación se mide por la precisión predictiva.

Ejemplos

- Predecir si un correo electrónico es spam o no spam.
- Predecir si un paciente es propenso a padecer enfermedades cardíacas.
- Estimar la tasa de ozono en una ciudad teniendo en cuenta variables climáticas.
- Predecir la ausencia o presencia de una especie en un entorno determinado.
- Predecir las fugas de clientes de una entidad financiera.
- Identificar cifras manuscritas de tarjetas postales en sobres.
- Dividir una población en varios subgrupos.

Modelización Estadística

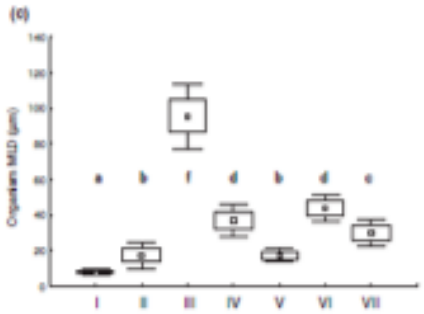


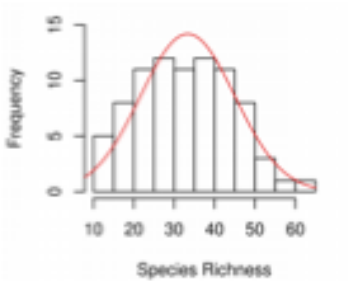
Modelización Estadística...

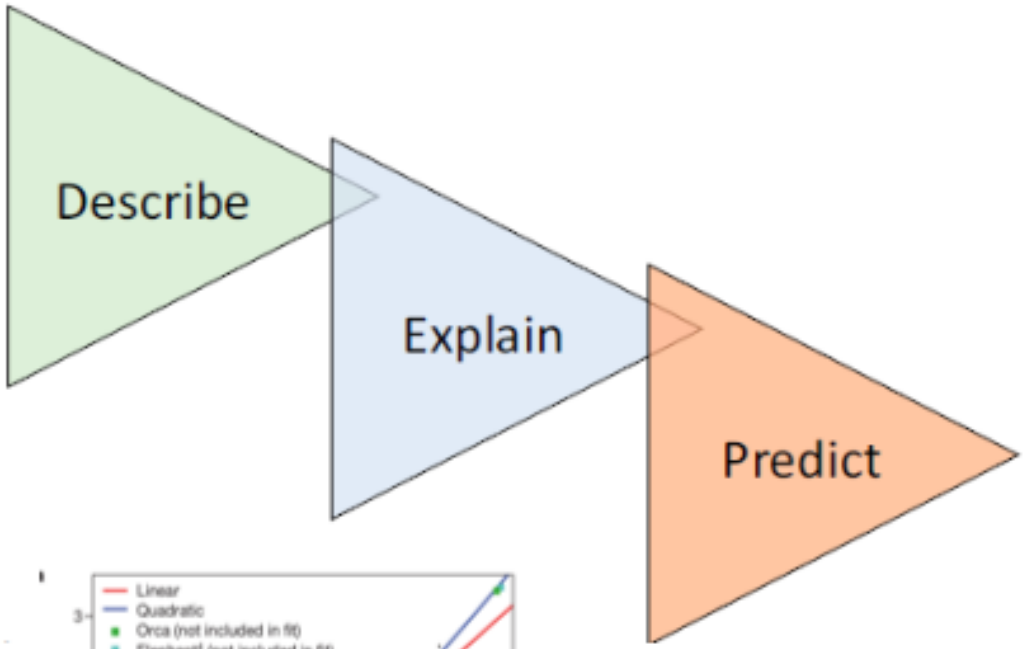
Cuando se está pensando en utilizar un modelo, es útil recordar:

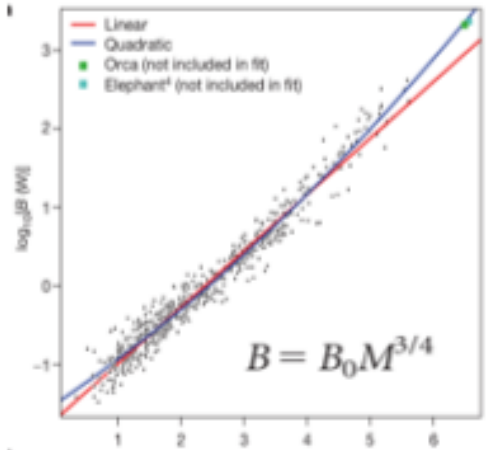
- la cultura de modelado de datos - cultura de modelado de algoritmos
- Supervisado - No supervisado
- Supervisado: Clasificación o Regresión
- Tipos de variables, Valores faltantes
- La precisión del método es importante, pero lo es aún más sobre unos *datos de prueba*.
- Multiplicidad de buenos modelos: métodos de agregación
- Dilemna de la navaja de Occam: ¿más simple es mejor? ¿Simplicidad frente a precisión?
- ¿Maldición de la dimensionalidad? ¿Impedimento o bendición?
- *(The focus)..is on solving the problem instead of asking what data model (they can create). The best solution could be an algorithmic model, or may be a data model, or may be a combination. But the trick to being a scientist is to be open to using a wide variety of tools*, Breiman, Las dos culturas.
- **Todos los modelos son erróneos, pero algunos son útiles**, Georges Box (1919-2013)

Recordar...









Kolokotrones et al 2010