Regresión

Tratamiento Inteligente de Datos
Master Universitario en Ingeniería Informática

UNIVERSIDAD DE GRANADA

Gabriel Navarro (*gnavarro@ugr.es*, *gnavarro@decsai.ugr.es*)

Objetivos

- Entender el problema de la regresión como técnica de clasificación de variables numéricas
- Conocer los ajustes usuales: lineal, polinómico, logarítmico, exponencial y logística
- Conocer las medidas más comunes de evaluación de la regresión
- Entender el concepto de árbol de regresión

Índice

- El problema de la regresión
- Regresión
  - Lineal simple y múltiple
  - Polinómica
  - Logarítmica
  - Exponencial
  - Logística
- Medidas de evaluación
- Árboles de regresión

El problema de la regresión

Las técnicas de clasificación del tema anterior predicen, para una instancia/caso/ejemplo, una clase entre un conjunto discreto de ellas

- Clasificamos según una variable nominal
- Puede ser numérica… pero se discretizaba/categorizaba

¿Y si queremos clasificar en una variable numérica real?

- Toma un número potencialmente infinito de valores
- Sin discretizar

El problema de la regresión

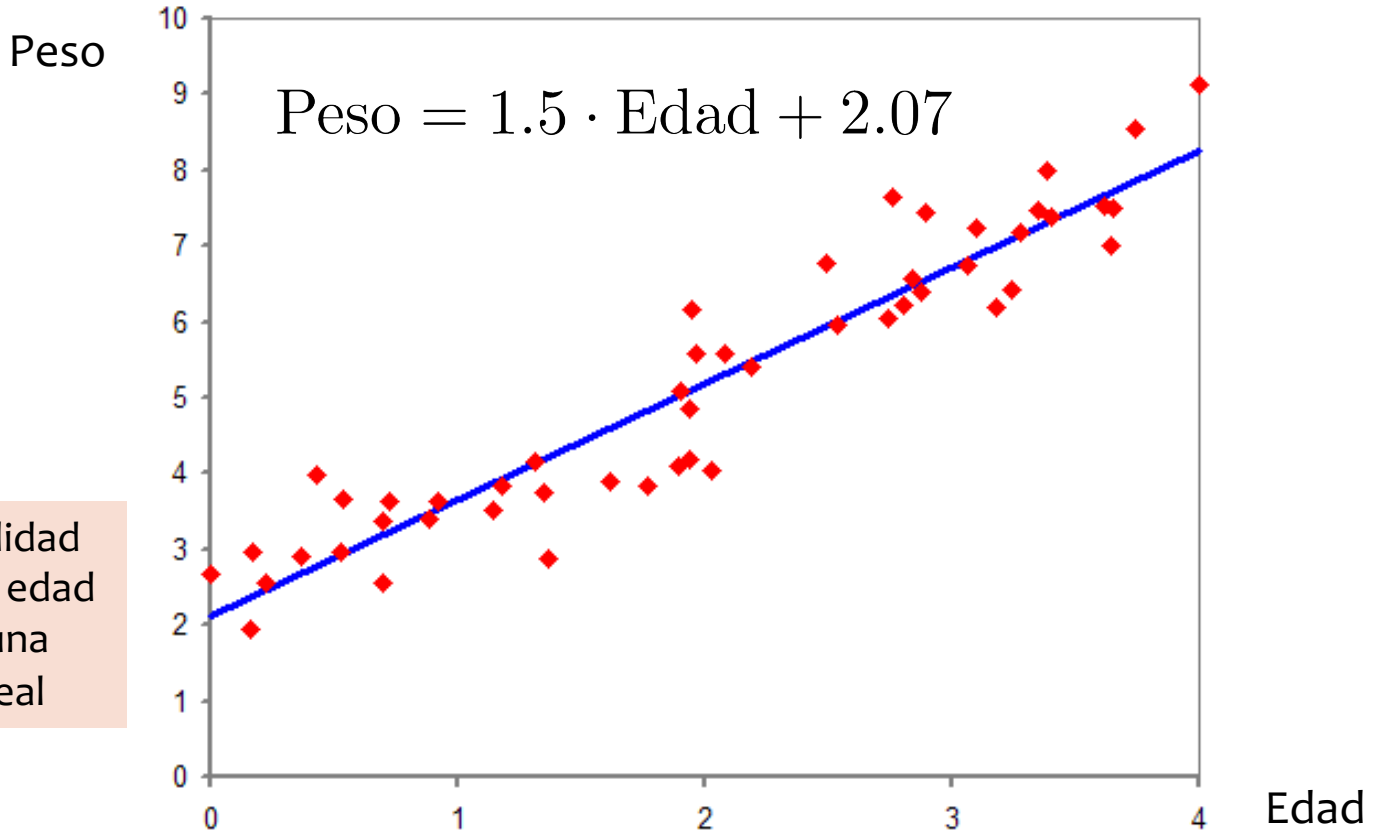
La regresión da una solución a este problema. La idea es similar a la clasificación

- Modelar la variable a clasificar con los datos disponibles
  - Normalmente, las variables son numéricas
  - Se modela la variable como función de las anteriores, es decir, la dependencia de la clase a clasificar respecto del resto de atributos
- Predecir nuevos datos a partir del modelo
  - Se substituye el valor de cada variable independiente
  - El valor con el que se clasifica es la imagen de la función

El problema de la regresión

Por ejemplo, teniendo en cuenta los datos de varios niños

¿Cuánto pesará mi bebé?

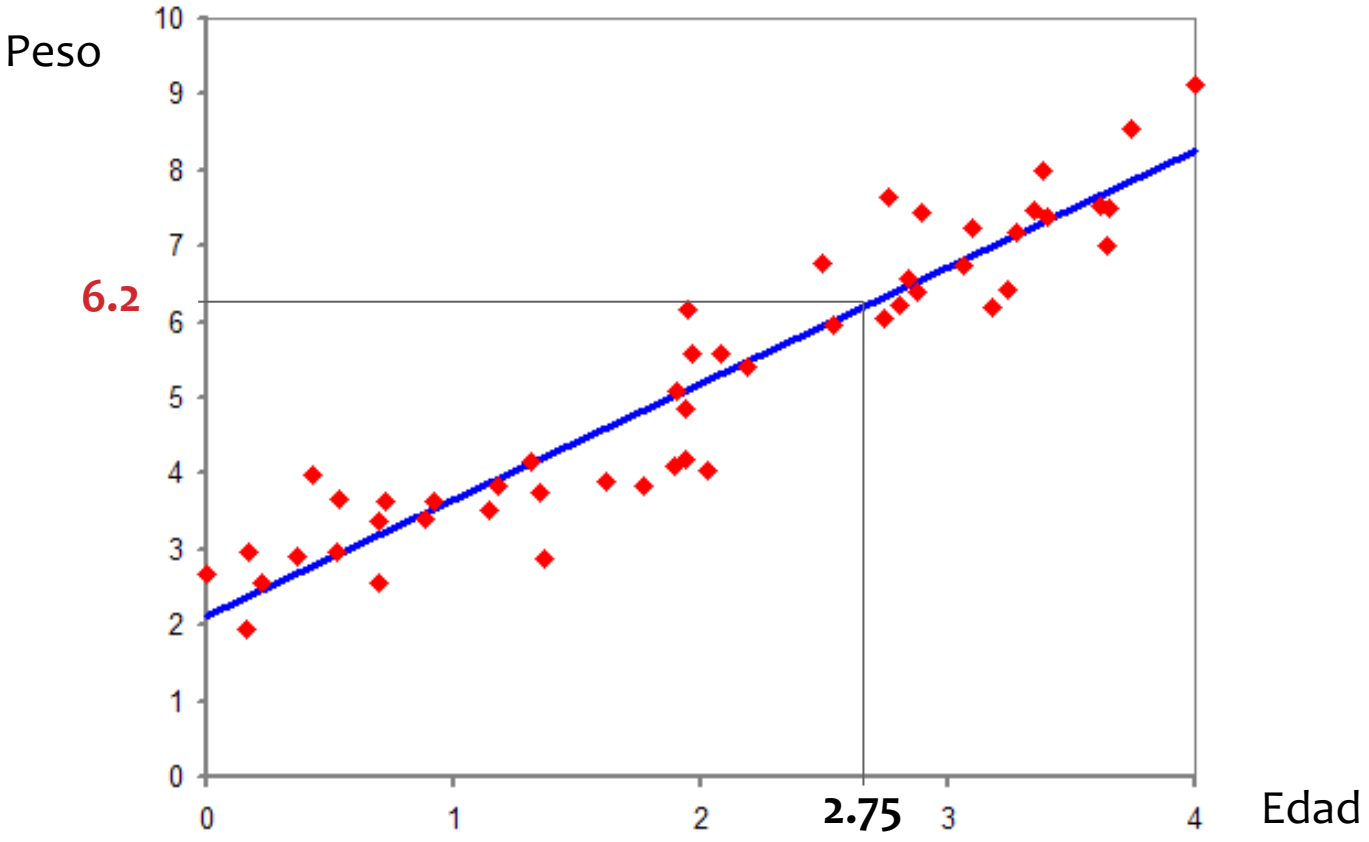


Ojo! En realidad El peso y la edad no siguen una relación lineal

El problema de la regresión

Por ejemplo, teniendo en cuenta los datos de varios niños

¿Cuánto pesará mi bebé?



El problema de la regresión

En general, expresamos la variable a clasificar en la forma

$$Y = f(X_1, X_2, \ldots, X_n) + \epsilon$$

donde

- f es la función sobre la que se realiza el ajuste
- Y es la variable dependiente o explicada
- $X_1, X_2, \ldots, X_n$ son las variables independientes o explicativas
- ϵ es una variable aleatoria que determina la tasa de error y se modela por una normal $\mathcal{N}(0, \sigma^2)$

El problema de la regresión

Tipos de regresión:

- Atendiendo al número de variables independientes
  - Simple, si sólo hay una
  - Múltiple, si hay más de una
- Atendiendo al tipo de función
  - Lineal
  - Polinomial
  - Logarítmica
  - Exponencial
  - Logística
  - ...

El problema de la regresión

Dos etapas a resolver:

- Elegir el tipo de función que modela la variable dependiente
  - No es nada fácil… tipos de funciones hay muchas!
  - Por la experiencia
  - Mediante una visualización de los datos
- Concretar la función estimando ciertos parámetros
  - Mediante una optimización de los parámetros
  - Los que minimizan el error cuadrático

El problema de la regresión

Etapa extra

Estimar la variable aleatoria que modela el error. Nos basamos en cuatro hipótesis:

- H1. El error sigue una normal
- H2. El error en un individuo es independiente del otro
- H3. El error en un individuo es independiente de las variables explicativas
- H4. El error en los individuos está generado por una variable aleatoria con la misma varianza

El problema de la regresión

Etapa extra

Estimar la variable aleatoria que modela el error. Nos basamos en cuatro hipótesis:

- H1. El error sigue una normal
- H2. El error en un individuo es independiente del otro
- H3. El error en un individuo es independiente de las variables explicativas
- H4. El error en los individuos está generado por una variable aleatoria con la misma varianza

$$\mathcal{N}(0, \sigma^2)$$

¿Cómo se calcula?

El problema de la regresión

Etapa extra

Estimar la variable aleatoria que modela el error

Calculamos la varianza utilizando el modelo de regresión obtenido

$$\widetilde{Y} = f(X_1, \ldots, X_n)$$

entonces

$$\sigma^2 = Var(\widetilde{Y} - Y)$$

Regresión lineal

La forma más sencilla de relacionar variables es mediante una expresión lineal

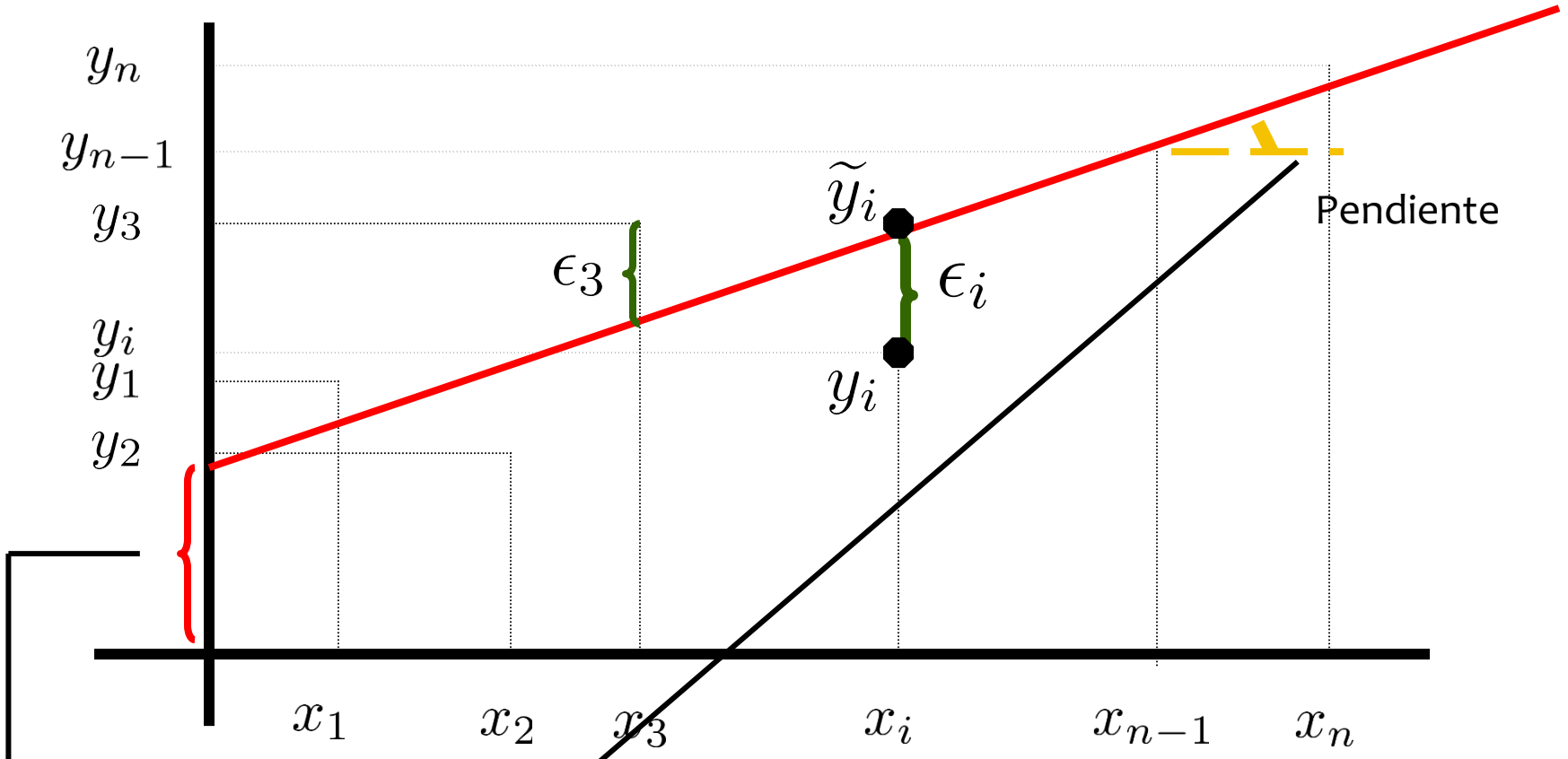
$$Y = aX + b$$ ← Recta de regresión

$$Y = a_1X_1 + a_2X_2 + b$$ ← Plano de regresión

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$$

Hiperplano de regresión

Regresión lineal simple



$$y_i = a + bx_i + \epsilon_i$$

Error

$$\epsilon_i = y_i - \widetilde{y}_i$$

Regresión lineal simple

La metodología para la obtención de la recta será hacer mínima la suma de los cuadrados del error

- ¿Por qué el cuadrado?

$$\min\left(\sum_{i=1}^{n}(\widetilde{y}_i - y_i)^2\right) = \min\left(\sum_{i=1}^{n}(a + bx_i - y_i)^2\right)$$

Regresión lineal simple

La metodología para la obtención de la recta será hacer mínima la suma de los cuadrados del error

- ¿Por qué el cuadrado?

$$\min\left(\sum_{i=1}^{n}(\widetilde{y}_i - y_i)^2\right) = \min\left(\sum_{i=1}^{n}(a + bx_i - y_i)^2\right)$$

$$\phi = \min\left(\sum_{i=1}^{n}(a + bx_i - y_i)^2\right)$$

derivamos respecto de los parámetros!

Regresión lineal simple

$$\frac{d\phi}{da} = 2\sum_{i=1}^{n}(a + bx_i - y_i) = an + b\sum x_i - \sum y_i = 0$$

Regresión lineal simple

$$\frac{d\phi}{da} = 2\sum_{i=1}^{n}(a + bx_i - y_i) = an + b\sum x_i - \sum y_i = 0$$

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} \Longrightarrow \boxed{a = \overline{y} - b\overline{x}}$$

Regresión lineal simple

$$\frac{d\phi}{da} = 2\sum_{i=1}^{n}(a + bx_i - y_i) = an + b\sum x_i - \sum y_i = 0$$

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} \Longrightarrow \boxed{a = \overline{y} - b\overline{x}}$$

$$\frac{d\phi}{db} = 2\sum_{i=1}^{n} x_i\,(a + bx_i - y_i) = 2\left(a\sum x_i + b\sum x_i^2 - \sum y_i x_i\right) = 0$$

Regresión lineal simple

$$\frac{d\phi}{da} = 2\sum_{i=1}^{n}(a + bx_i - y_i) = an + b\sum x_i - \sum y_i = 0$$

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} \Longrightarrow \boxed{a = \overline{y} - b\overline{x}}$$

$$\frac{d\phi}{db} = 2\sum_{i=1}^{n} x_i\,(a + bx_i - y_i) = 2\left(a\sum x_i + b\sum x_i^2 - \sum y_i x_i\right) = 0$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 = \overline{y}n\overline{x} - bn\overline{x}\overline{x} + b\sum x^2$$

Regresión lineal simple

$$\frac{d\phi}{da} = 2\sum_{i=1}^{n}(a + bx_i - y_i) = an + b\sum x_i - \sum y_i = 0$$

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} \Longrightarrow \boxed{a = \overline{y} - b\overline{x}}$$

$$\frac{d\phi}{db} = 2\sum_{i=1}^{n} x_i\,(a + bx_i - y_i) = 2\left(a\sum x_i + b\sum x_i^2 - \sum y_i x_i\right) = 0$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 = \overline{y}n\overline{x} - bn\overline{x}\overline{x} + b\sum x^2$$

$$\sum (x_i y_i - \overline{xy}) = b\sum \left(x^2 - \overline{x}^2\right)$$

Es una forma de poner la covarianza

Es una forma de poner la varianza

Regresión lineal simple

$$\frac{d\phi}{da} = 2\sum_{i=1}^{n}(a + bx_i - y_i) = an + b\sum x_i - \sum y_i = 0$$

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} \Longrightarrow \boxed{a = \overline{y} - b\overline{x}}$$

$$\frac{d\phi}{db} = 2\sum_{i=1}^{n} x_i\,(a + bx_i - y_i) = 2\left(a\sum x_i + b\sum x_i^2 - \sum y_i x_i\right) = 0$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 = \overline{y}n\overline{x} - bn\overline{x}\overline{x} + b\sum x^2$$

$$\sum (x_i y_i - \overline{xy}) = b\sum \left(x^2 - \overline{x}^2\right)$$

$$n\sigma_{XY} = bn\sigma_X^2 \Longrightarrow \boxed{b = \frac{\sigma_{XY}}{\sigma_X^2}}$$

Regresión lineal simple

Realizar un ajuste lineal para los datos

Altura (A)	1.4	1.56	1.67	1.63	1.8	1.85	1.69
Peso (P)	34	45	56	55	70	76	65

Regresión lineal simple

								Suma	Media
Altura (A)	1,4	1,56	1,67	1,63	1,8	1,85	1,69	11,6	1,66
Peso (P)	34	45	56	55	70	76	65	401	57,29
A^2	1,96	2,43	2,79	2,66	3,24	3,42	2,86		
$A^2-(\overline{A})^2$	-0,79	-0,31	0,04	-0,09	0,49	0,68	0,11	0,14	0,02
$\overline{AP}$	47,6	70,2	93,52	89,65	126	140,6	109,85		
AP-AP	-47,33	-24,73	-1,41	-5,28	31,07	45,67	14,92	12,91	1,84

Regresión lineal simple

								Suma	Media
Altura (A)	1,4	1,56	1,67	1,63	1,8	1,85	1,69	11,6	1,66
Peso (P)	34	45	56	55	70	76	65	401	57,29
A^2	1,96	2,43	2,79	2,66	3,24	3,42	2,86		
$A^2-(\overline{A})^2$	-0,79	-0,31	0,04	-0,09	0,49	0,68	0,11	0,14	**0,02**
AP	47,6	70,2	93,52	89,65	126	140,6	109,85		
$AP-\overline{AP}$	-47,33	-24,73	-1,41	-5,28	31,07	45,67	14,92	12,91	**1,84**

$$b = \frac{\sigma_{AP}}{\sigma_A^2} = \frac{1.84}{0.02} = 92.5$$

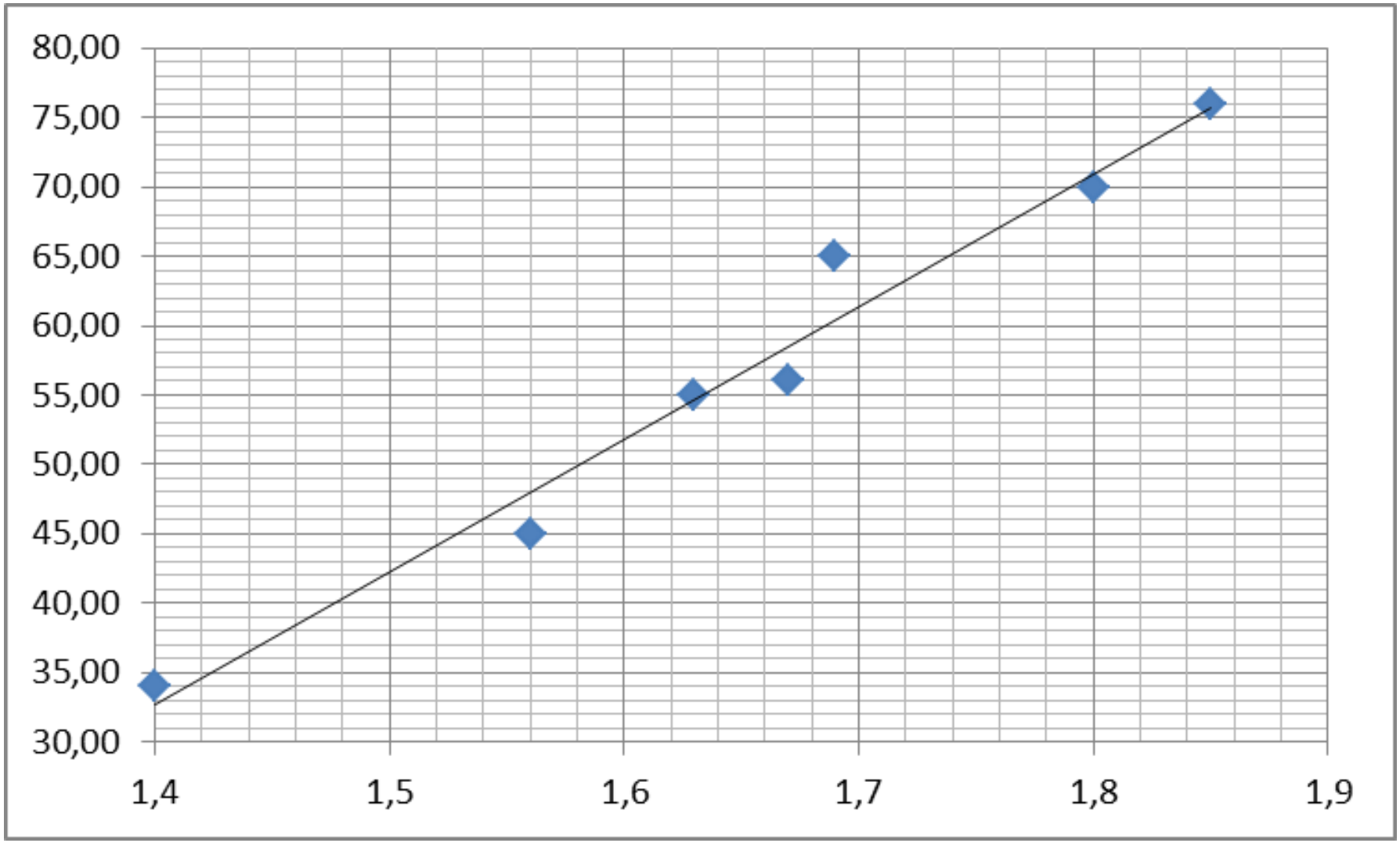
Regresión lineal simple

								Suma	Media
Altura (A)	1,4	1,56	1,67	1,63	1,8	1,85	1,69	11,6	**1,66**
Peso (P)	34	45	56	55	70	76	65	401	**57,29**
A^2	1,96	2,43	2,79	2,66	3,24	3,42	2,86		
$A^2-(\overline{A})^2$	-0,79	-0,31	0,04	-0,09	0,49	0,68	0,11	0,14	0,02
AP	47,6	70,2	93,52	89,65	126	140,6	109,85		
$AP-\overline{AP}$	-47,33	-24,73	-1,41	-5,28	31,07	45,67	14,92	12,91	1,84

$$b = \frac{\sigma_{AP}}{\sigma_A^2} = \frac{1.84}{0.02} = 92.5$$

$$a = \overline{P} - b\overline{A} = 57.29 - 92.5 \cdot 1.66 = -96.26$$

Regresión lineal simple



Regresión lineal simple

								Suma	Media
Altura (A)	1,4	1,56	1,67	1,63	1,8	1,85	1,69	11,6	1,66
Peso (P)	34	45	56	55	70	76	65	401	57,29
Peso Est. (PE)	33,24	48,04	58,22	54,52	70,24	74,87	60,07		
P-PE	0,76	-3,04	-2,21	0,49	-0,24	1,14	4,94		
$(P-PE)^2$	0,58	9,24	4,91	0,24	0,06	1,29	24,35	40,66	5,81

$$\widetilde{P} = -96.26 + 92.5A$$

Regresión lineal simple

								Suma	Media
Altura (A)	1,4	1,56	1,67	1,63	1,8	1,85	1,69	11,6	1,66
Peso (P)	34	45	56	55	70	76	65	401	57,29
Peso Est. (PE)	33,24	48,04	58,22	54,52	70,24	74,87	60,07		
P-PE	0,76	-3,04	-2,21	0,49	-0,24	1,14	4,94		
$(P-PE)^2$	0,58	9,24	4,91	0,24	0,06	1,29	24,35	40,66	**5,81**

$$\widetilde{P} = -96.26 + 92.5A$$

$$P \approx -96.26 + 92.5A + \mathcal{N}(0, 5.81)$$

Regresión lineal múltiple

En este caso, existe más de una variable explicativa, y ajustamos según una ecuación lineal del tipo

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b + \epsilon$$

minimizando el error al cuadrado

$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} (b + a_1 x_1 + \ldots + a_n x_n - y_i)^2 \right)$$

Regresión lineal múltiple

En este caso, existe más de una variable explicativa, y ajustamos según una ecuación lineal del tipo

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b + \epsilon$$

minimizando el error al cuadrado

$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} (b + a_1 x_1 + \ldots + a_n x_n - y_i)^2 \right)$$

Derivamos sobre los parámetros

Regresión lineal múltiple

$$
\begin{cases}
\frac{d\phi}{db} = 2\sum_{i=1}^{m}(b + a_1x_{1i} + \ldots + a_nx_{ni} - y_i) = 0 \\
\frac{d\phi}{da_1} = 2\sum_{i=1}^{m} x_{1i}(b + a_1x_{1i} + \ldots + a_nx_{ni} - y_i) = 0 \\
\vdots \\
\frac{d\phi}{da_n} = 2\sum_{i=1}^{m} x_{ni}(b + a_1x_{1i} + \ldots + a_nx_{ni} - y_i) = 0
\end{cases}
$$

$$
\begin{cases}
bm + a_1\sum x_{1i} + \ldots + a_n\sum x_{ni} = \sum y_i \\
b\sum x_{1i} + a_1\sum x_{1i}^2 + \ldots + a_n\sum x_{1i}x_{ni} = \sum x_{1i}y_i \\
\vdots \\
b\sum x_{ni} + a_1\sum x_{ni}x_{1i} + \ldots + a_n\sum x_{ni}^2 = \sum x_{ni}y_i
\end{cases}
$$

Regresión lineal múltiple

$$\underbrace{\begin{pmatrix} m & \sum x_{1i} & \cdots & \sum x_{ni} \\ \sum x_{1i} & \sum x_{1i}^2 & \cdots & \sum x_{1i}x_{ni} \\ \vdots & & & \vdots \\ \sum x_{ni} & \sum x_{ni}x_{1i} & \cdots & \sum x_{ni}^2 \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} b \\ a_1 \\ \vdots \\ a_n \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \vdots \\ \sum x_{ni}y_i \end{pmatrix}}_{N}$$

Los parámetros se obtienen al resolver el sistema de ecuaciones

$$M \cdot A = N$$

Regresión polinómica

Se trata de ajustar los datos por una función del tipo

$$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n + \epsilon$$

minimizando el error al cuadrado

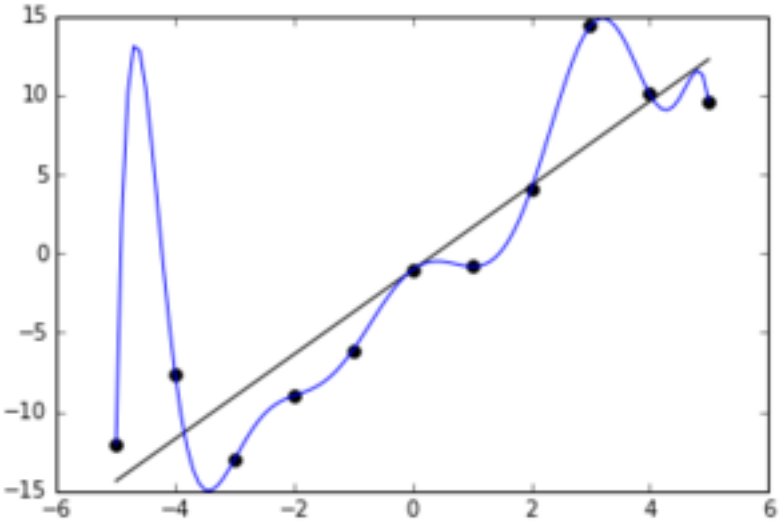
$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} (a_0 + a_1 x + \ldots + a_n x^n - y_i)^2 \right)$$

Regresión polinómica

Se trata de ajustar los datos por una función del tipo

$$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n + \epsilon$$

minimizando el error al cuadrado



Es importante ajustar el valor de n, ya que, dados m puntos, se puede encontrar una curva de grado m – 1 que pase por todos ellos (sobreaprendizaje)

Regresión polinómica

Se trata de ajustar los datos por una función del tipo

$$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n + \epsilon$$

minimizando el error al cuadrado

$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} (a_0 + a_1 x + \ldots + a_n x^n - y_i)^2 \right)$$

Derivamos sobre los parámetros

Regresión polinómica

$$
\begin{cases}
\frac{d\phi}{da_0} = & 2\sum_{i=1}^{n}\left(a_0 + a_1 x_i + \ldots + a_n x_i^n - y_i\right) = 0 \\
\frac{d\phi}{da_1} = & 2\sum_{i=1}^{n} x_i\left(a_0 + a_1 x_i + \ldots + a_n x_i^n - y_i\right) = 0 \\
\vdots & \\
\frac{d\phi}{da_n} = & 2\sum_{i=1}^{n} x_i^n\left(a_0 + a_1 x_i + \ldots + a_n x_i^n - y_i\right) = 0
\end{cases}
$$

Regresión polinómica

$$
\begin{cases}
\frac{d\phi}{da_0} = & 2\sum_{i=1}^{n}(a_0 + a_1x_i + \ldots + a_nx_i^n - y_i) = 0 \\
\frac{d\phi}{da_1} = & 2\sum_{i=1}^{n} x_i\,(a_0 + a_1x_i + \ldots + a_nx_i^n - y_i) = 0 \\
\vdots & \\
\frac{d\phi}{da_n} = & 2\sum_{i=1}^{n} x_i^n\,(a_0 + a_1x_i + \ldots + a_nx_i^n - y_i) = 0
\end{cases}
$$

$$
\begin{cases}
a_0 n & +a_1\sum x_i & +\ldots & +a_n\sum x_i^n & = \sum y_i \\
a_0\sum x_i & +a_1\sum x_i^2 & +\ldots & +a_n\sum x_i^{n+1} & = \sum x_i y_i \\
 & & \vdots & & \\
a_0\sum x_i^n & +a_1\sum x_i^{n+1} & +\ldots & +a_n\sum x_i^{2n} & = \sum x_i^n y_i
\end{cases}
$$

Regresión polinómica

$$\begin{pmatrix} n & \sum x_i & \cdots & \sum x_i^n \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{n+1} \\ \vdots & & & \vdots \\ \sum x_i^n & \sum x_i^{n+1} & \cdots & \sum x_i^{2n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^n y_i \end{pmatrix}$$

Regresión polinómica

$$\underbrace{\begin{pmatrix} n & \sum x_i & \cdots & \sum x_i^n \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{n+1} \\ \vdots & & & \vdots \\ \sum x_i^n & \sum x_i^{n+1} & \cdots & \sum x_i^{2n} \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^n y_i \end{pmatrix}}_{N}$$

Los parámetros se obtienen al resolver el sistema de ecuaciones

$$M \cdot A = N$$

Regresión polinómica

Realizar un ajuste cuadrático para los datos

X	Y
1	1.25
2	5
3	11.25
4	20
5	30.5

Regresión polinómica

	X	Y	X^2	X^3	X^4	XY	X^2Y
	1	1.25	1	1	1	1.25	1.25
	2	5	4	8	16	10	20
	3	11.25	9	27	81	33.75	101.5
	4	20	16	64	256	80	320
	5	30.5	25	125	625	152.5	762.5
Suma	15	68	55	225	979	277.5	1205

Regresión polinómica

	X	Y	X^2	X^3	X^4	XY	X^2Y
	1	1.25	1	1	1	1.25	1.25
	2	5	4	8	16	10	20
	3	11.25	9	27	81	33.75	101.5
	4	20	16	64	256	80	320
	5	30.5	25	125	625	152.5	762.5
Suma	15	68	55	225	979	277.5	1205

$$\begin{pmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 68 \\ 277.5 \\ 1205 \end{pmatrix}$$

Regresión polinómica

	X	Y	X^2	X^3	X^4	XY	X^2Y
	1	1.25	1	1	1	1.25	1.25
	2	5	4	8	16	10	20
	3	11.25	9	27	81	33.75	101.5
	4	20	16	64	256	80	320
	5	30.5	25	125	625	152.5	762.5
Suma	15	68	55	225	979	277.5	1205

$$\begin{pmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 68 \\ 277.5 \\ 1205 \end{pmatrix}$$

$$(-0.945412844036703, 3.12866972477065, 0.469036697247706)$$

Regresión polinómica

	X	Y	X^2	X^3	X^4	XY	X^2Y	$\tilde{Y}$	$\in=Y-\tilde{Y}$	$\in^2$
	1	1.25	1	1	1	1.25	1.25	2.65	-1.4	1.96
	2	5	4	8	16	10	20	7.19	-2.19	4.8
	3	11.25	9	27	81	33.75	101.5	12.67	-1.42	2.02
	4	20	16	64	256	80	320	19.09	0.91	0.83
	5	30.5	25	125	625	152.5	762.5	26.45	4.05	16.04
Suma	15	68	55	225	979	277.5	1205	68.08	**0.05**	25.65

$$\tilde{Y} = -0.95 + 3.13X + 0.47X^2$$

No sale cero!!!
¿Por qué?

Regresión polinómica

	X	Y	X^2	X^3	X^4	XY	X^2Y	$\tilde{Y}$	$\in=Y-\tilde{Y}$	$\in^2$
	1	1.25	1	1	1	1.25	1.25	2.65	-1.4	1.96
	2	5	4	8	16	10	20	7.19	-2.19	4.8
	3	11.25	9	27	81	33.75	101.5	12.67	-1.42	2.02
	4	20	16	64	256	80	320	19.09	0.91	0.83
	5	30.5	25	125	625	152.5	762.5	26.45	4.05	16.04
Suma	15	68	55	225	979	277.5	1205	68.08	**0.05**	25.65

$$\tilde{Y} = -0.95 + 3.13X + 0.47X^2$$

$$Y \approx -0.95 + 3.13X + 0.47X^2 + \mathcal{N}(0, 8.55)$$

No sale cero!!!
¿Por qué?

Regresión logarítmica

Se trata de ajustar los datos a una función del tipo

$$Y = a \ln X + b + \epsilon$$

minimizando el error al cuadrado

$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} (a \ln x_i + b - y_i)^2 \right)$$

Regresión logarítmica

Se trata de ajustar los datos a una función del tipo

$$Y = a \ln X + b + \epsilon$$

minimizando el error al cuadrado

$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} (a \ln x_i + b - y_i)^2 \right)$$

La solución es la misma que la regresión lineal pero considerando la variable $X' = \ln X$

Regresión logarítmica

Realiza un ajuste logarítmico para los datos

X	Y
1	1.25
2	5
3	11.25
4	20
5	30.5

Regresión logarítmica

	X	Y	ln X	$(\ln X)^2$	Y ln X	Y ln X-$\overline{Y}$ $\overline{\ln X}$	$(\ln X)^2$-$(\overline{\ln X})^2$
	1	1.25	0	0	0	-13.06	-0.92
	2	5	0.69	0.48	3.45	-9.61	-0.44
	3	11.25	1.1	1,21	12.38	-0.69	0.29
	4	20	1.39	1.93	27.8	14,74	1.01
	5	30.5	1.61	2.59	49.1	36,05	1.67
Suma		68	4.79			27,43	1.61
Media		13,6	0.96			5,49	0.32

$$b = \frac{\sigma_{\ln X,Y}}{\sigma^2_{\ln X}} = \frac{5,49}{0.32} = 17.16$$

$$a = \overline{\ln X} - b\overline{Y} = 0.96 - 17.16 \cdot 13.6 = -232.41$$

Regresión logarítmica

	X	Y	$\widetilde{Y}$	Y-$\widetilde{Y}$
	1	1.25	17.6	-16.35
	2	5	-143.5	148.5
	3	11.25	-237.73	248.98
	4	20	-304.59	324.59
	5	30.5	-356.45	386.95
Suma		68		1092.67
Media		13,6		**156.1**

No parece muy buen ajuste!

$$\widetilde{Y} = -232.41 \ln X + 17.6$$

Regresión logarítmica

	X	Y	$\widetilde{Y}$	Y-$\widetilde{Y}$
	1	1.25	17.6	-16.35
	2	5	-143.5	148.5
	3	11.25	-237.73	248.98
	4	20	-304.59	324.59
	5	30.5	-356.45	386.95
Suma		68		1092.67
Media		13,6		**156.1**

$$\widetilde{Y} = -232.41 \ln X + 17.6$$

$$Y \approx -232.41 \ln X + 17.6 + \mathcal{N}(0, 156.1)$$

Regresión exponencial

Se trata de ajustar los datos a una función del tipo

$$Y = a \cdot b^X + \epsilon$$

minimizando el error al cuadrado

$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} a b^{x_i} - y_i \right)$$

Regresión exponencial

Se trata de ajustar los datos a una función del tipo

$$Y = a \cdot b^X + \epsilon$$

minimizando el error al cuadrado

$$\phi = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \left(\sum_{i=1}^{n} a b^{x_i} - y_i \right)$$

Problema: No es sencillo resolver el sistema

$$\frac{d\phi}{da} = 0 \quad \frac{d\phi}{db} = 0$$

Regresión exponencial

Solución

1. Tomar logaritmos en ambos lados

$$\ln Y = \ln a + b \ln X + \ln \epsilon$$

2. Substituir las variables

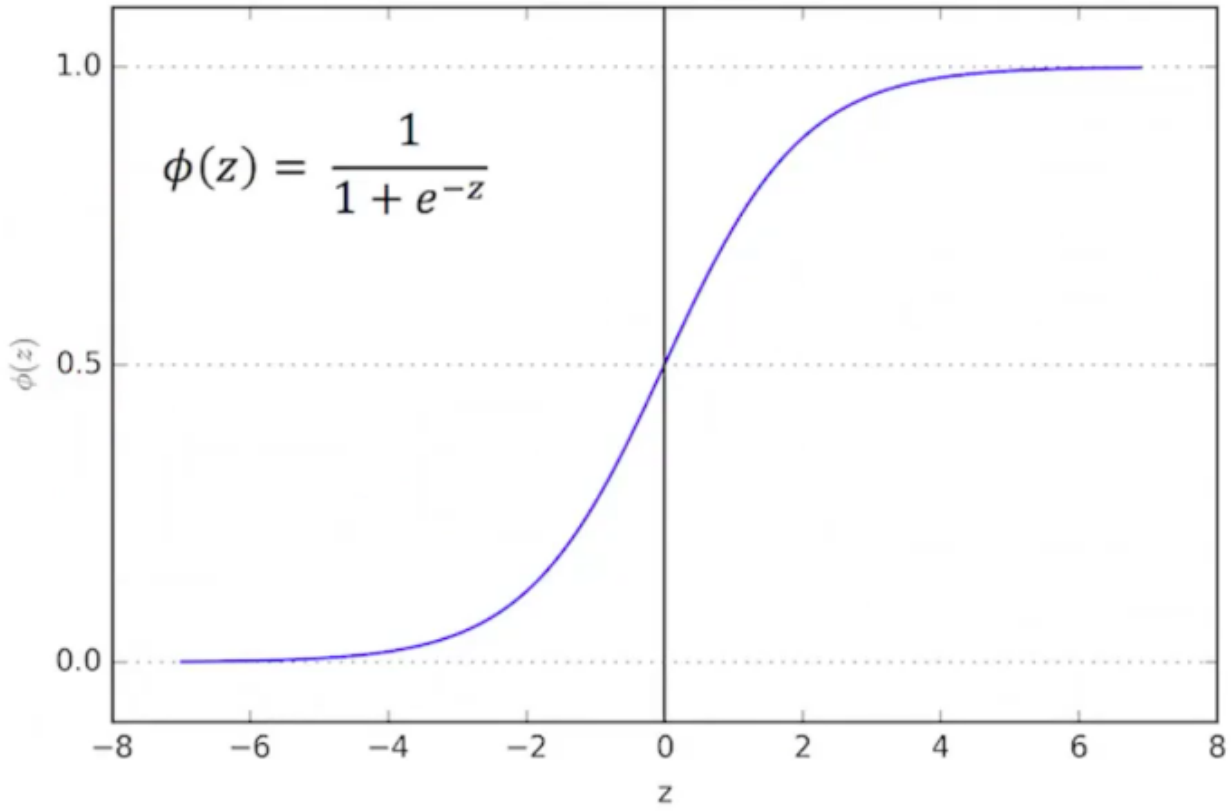
$$\left.\begin{aligned} Y' &= \ln Y \\ X' &= \ln X \\ a' &= \ln a \\ \epsilon' &= \ln \epsilon \end{aligned}\right\} \quad Y' = a' + bX' + \epsilon'$$

Lineal

3. Resolver para obtener los parámetros
4. Invertir el cambio, es decir, $a = e^{a'}$

Regresión logística

La regresión logística se usa para predecir la probabilidad de ocurrencia de una variable binaria (fallo, no-fallo; pago, impago;…).



Regresión logística

Supongamos que *p* es la probabilidad de que una instancia de la base de datos se clasifique como positivo y *1-p*, que clasifique como negativo

El modelo de regresión logística trata de ajustar linealmente

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

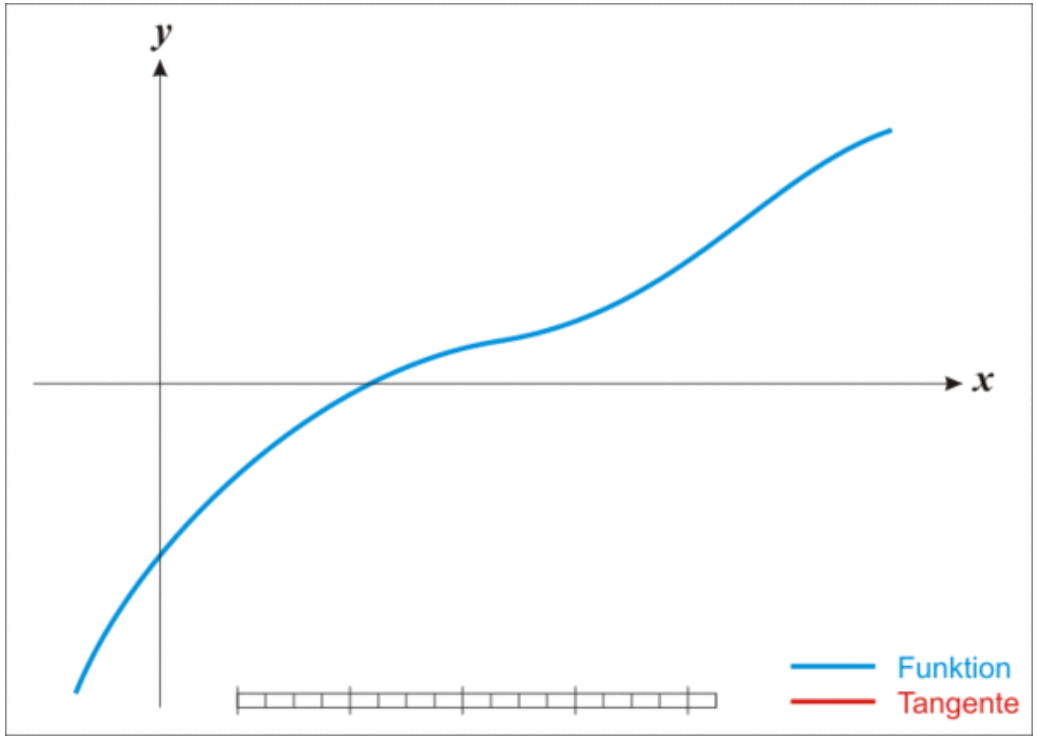
Tras unas pocas manipulaciones

$$p = \frac{1}{1 + e^{-(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Regresión logística

La estimación de los parámetros no se suele realizar utilizando mínimos cuadrados, si no con máxima verosimilitud (Máximum Likehood Estimation, MLE). Las ecuaciones que se obtienen no son lineales (como en los modelos de regresión anteriores) por lo que se utilizan métodos iterativos para calcular los coeficientes.

Newton-Raphson



Trabajo Evaluable

Realizar un trabajo sobre el modelo de regresión logística explicando en detalle la regresión logística, cómo se pueden estimar los parámetros, la medidas más usuales de la bondad del ajuste, extensión a y dando un pequeño ejemplo concreto de aplicación. Para tres personas. Tiempo de la exposición, 45 minutos.

Posible bibliografía: Hablar con el profesor

Medidas de evaluación

- Todas las metodologías de evaluación para clasificación son válidas para regresión

- Sin embargo, las métricas no son reutilizables
  - No podemos contar cuándo se acierta, "siempre" se falla…
  - La cuestión es "cuánto de cerca" me he quedado entre el valor real y el estimado

Medidas de evaluación

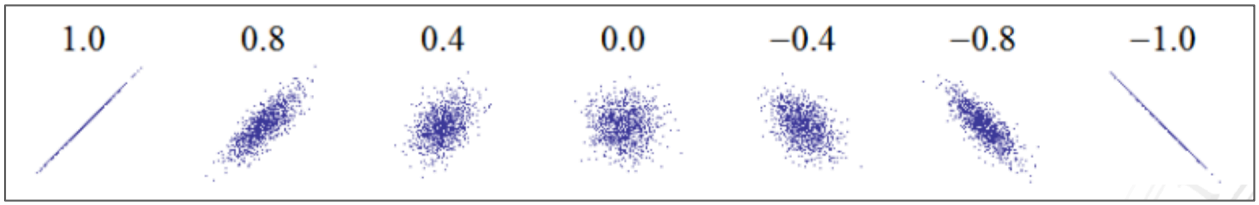
- Error cuadrático medio $$\mathrm{ECM} = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}_i - y_i)^2$$
- ECM estandarizado $$\mathrm{ECME} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}_i - y_i)^2}$$
- Error medio absoluto $$\mathrm{EMA} = \frac{1}{n}\sum_{i=1}^{n}|\widetilde{y}_i - y_i|$$
- Error absoluto relativo $$\mathrm{EAR} = \frac{1}{\sum_{i=1}^{n}|\widetilde{y}_i - \overline{y}|}\sum_{i=1}^{n}|\widetilde{y}_i - y_i|$$

Medidas de evaluación

- Coeficiente de correlación simple. Mide la relación lineal entre dos variables

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \sum_{i=1}^{n}(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Toma valores en [-1,1]



Medidas de evaluación

- Coeficiente de determinación

$$R^2 = \frac{\sum_{i=1}^{n} (\widetilde{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

  - Entre 0 y 1
  - 1 si hay correlación perfecta, toda la variación es explicada
  - 0 si no hay correlación

- Coeficiente de determinación ajustado. Tiene en cuenta la complejidad del modelo (el número de variables)

$$R^2_{aj} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

El proceso de regresión lineal

Cuando entre las variables independientes hay una relación lineal aparece la multicolinealidad. Se debe suprimir una de las variables relacionadas. Para hacer un análisis de regresión correcto:

- Análisis exploratorio: utilizar diagramas de dispersión (*scatter plots*) para visualizar una cierta dependencia lineal parcial entre la variable dependiente y las independientes
- Buscar posibles relaciones de multicolinealidad entre las variables independientes, $R^2 \geq 0.65$ es un buen límite. Eliminar variables relacionadas dejando sólo una de ellas
- Realizar el análisis de regresión, calculando coeficientes de regresión y determinación nuevamente
- 0.65, o más, es un buen valor de ajuste

Árboles de regresión

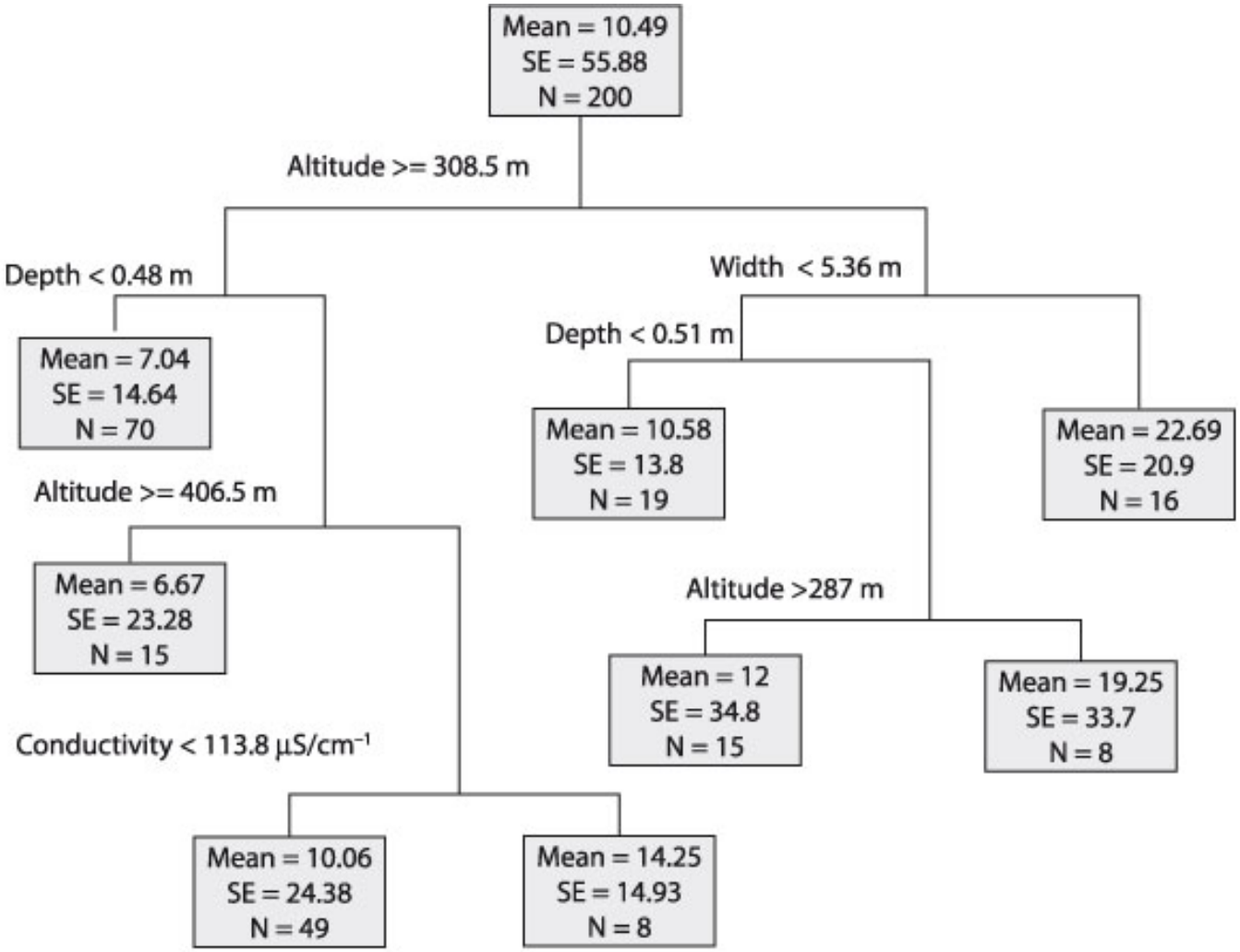
La idea de un árbol de regresión es la misma que la de un árbol de clasificación

- Sólo que la variable clase es numérica y continua (se discretiza)
- El valor final es la media de los ejemplos de esa hoja para la variable clase
- Se suele podar para evitar sobreentrenamiento
- La elección de cada variable: la que minimice la varianza de la variable objetivo

$$Var(N) - \sum_i \frac{\#N_i}{\#N} Var(N_i)$$

donde N es el conjunto de ejemplos en el nodo y N_i son los nodos hijos de N

Árboles de regresión



Árboles de regresión

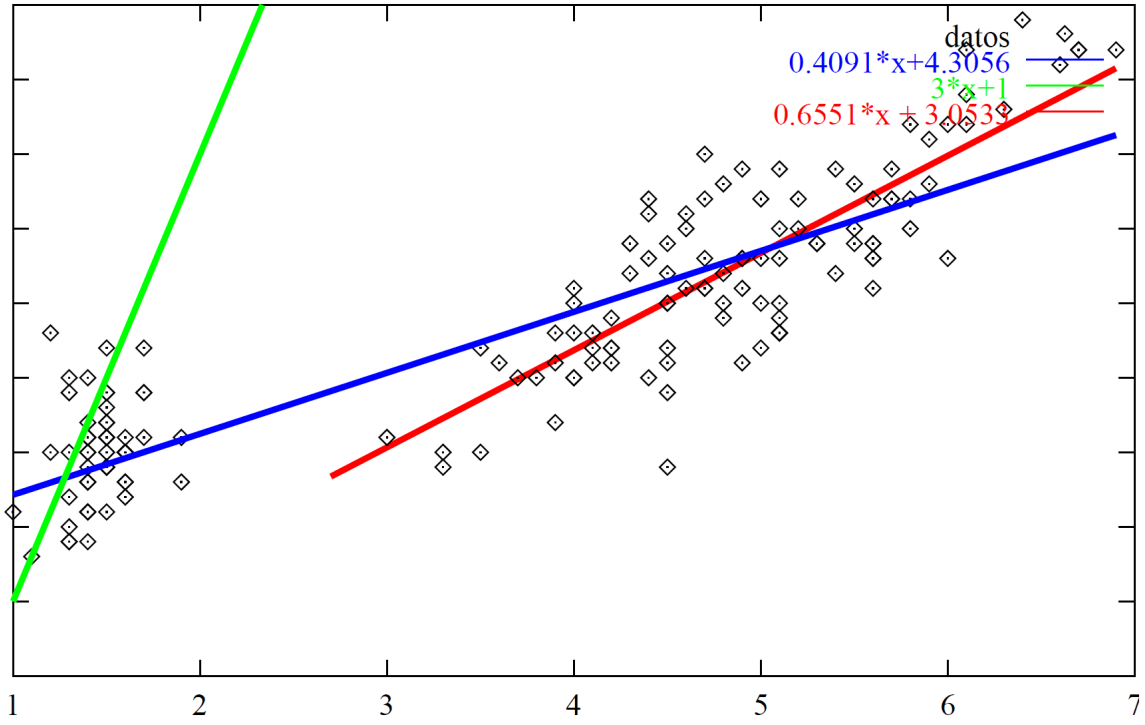
En general, al igual que en un árbol de decisión, hay que decidir varios criterios:

- ¿Cómo discretizar las variables explicativas?
- ¿Cuándo un nodo es un nodo hoja (cuándo se para)?
- ¿Qué valor se toma en un nodo hoja?

Por ejemplo, AID es como el ID3 pero con la escisión por varianza

Árboles de modelos

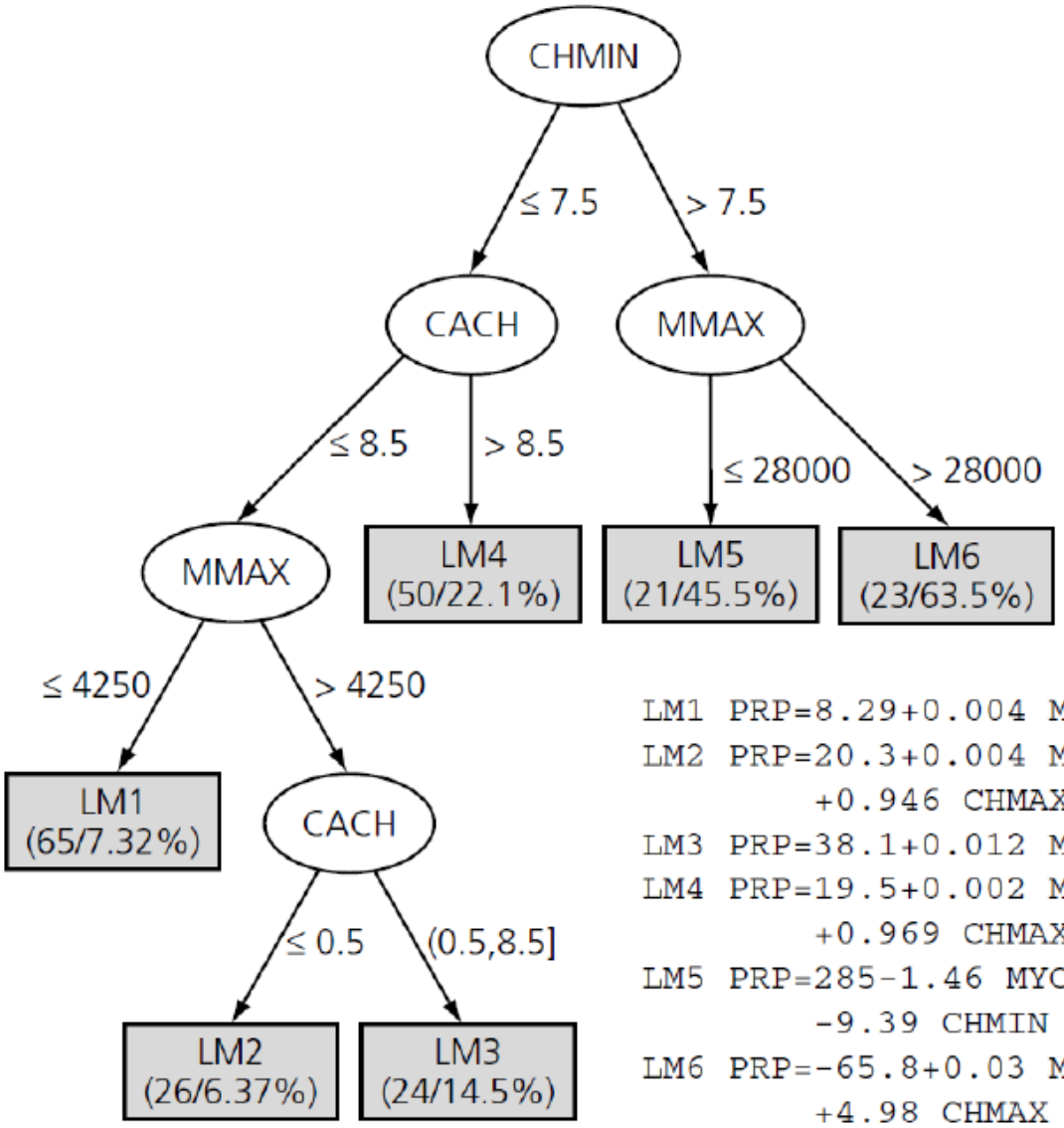
Consiste en crear un árbol en el que las hojas son funciones de regresión en vez de un valor



Árboles de modelos

	Cycle time (ns) MYCT	Main memory (KB) Min. MMIN	Main memory (KB) Max. MMAX	Cache (KB) CACH	Channels Min. CHMIN	Channels Max. CHMAX	Performance PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
3	29	8000	32000	32	8	32	220
4	29	8000	32000	32	8	32	172
5	29	8000	16000	32	8	16	132
. . .							
207	125	2000	8000	0	2	14	52
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Árboles de modelos




```
LM1  PRP=8.29+0.004 MMAX+2.77 CHMIN
LM2  PRP=20.3+0.004 MMIN-3.99 CHMIN
          +0.946 CHMAX
LM3  PRP=38.1+0.012 MMIN
LM4  PRP=19.5+0.002 MMAX+0.698 CACH
          +0.969 CHMAX
LM5  PRP=285-1.46 MYCT+1.02 CACH
          -9.39 CHMIN
LM6  PRP=-65.8+0.03 MMIN-2.94 CHMIN
          +4.98 CHMAX
```

Bibliografía

- Introducción a la Minería de Datos. José Hernández Orallo, M.José Ramírez Quintana, Cèsar Ferri Ramírez. Pearson, 2004. **Capítulo 7.**
- *M. R. Spiegel, J. Schiller, R. A. Srinivasan, Probabilidad y Estadística, 2° Edición, McGraw-Hill, 2003.* ***Capitulo 8.***