DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Regresión

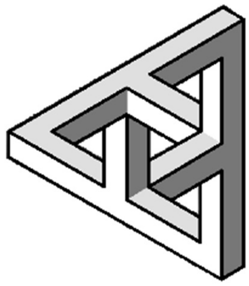
© Fernando Berzal, berzal@acm.org

Regresión

- Predicción: Análisis de regresión
- Regresión lineal
  - Regresión lineal simple
  - Coeficiente de correlación lineal
  - Regresión lineal múltiple
  - Gradiente descendente & Ecuación normal
- Regresión no lineal
  - Regresión logística
- Regularización
- Regresión local
  - LOESS
  - CART

Introducción

Regresión



Modelo
Predictivo y, en ocasiones, descriptivo.

Objetivo
Aprender una función que relaciona
un conjunto de variables independientes (X)
con una variable cuantitativa (y).

Datos
Numéricos.

Variantes y extensiones
Predicción de series temporales
[time series forecasting]

Predicción

La predicción (numérica) es…

Similar a la clasificación:

- Se construye un modelo a partir de un conjunto de entrenamiento.
- Se utiliza el modelo para predecir el valor de una variable (continua u ordenada).

Diferente a la clasificación:

- El modelo define una función continua.

Método más empleado: **Análisis de regresión**

Análisis de regresión

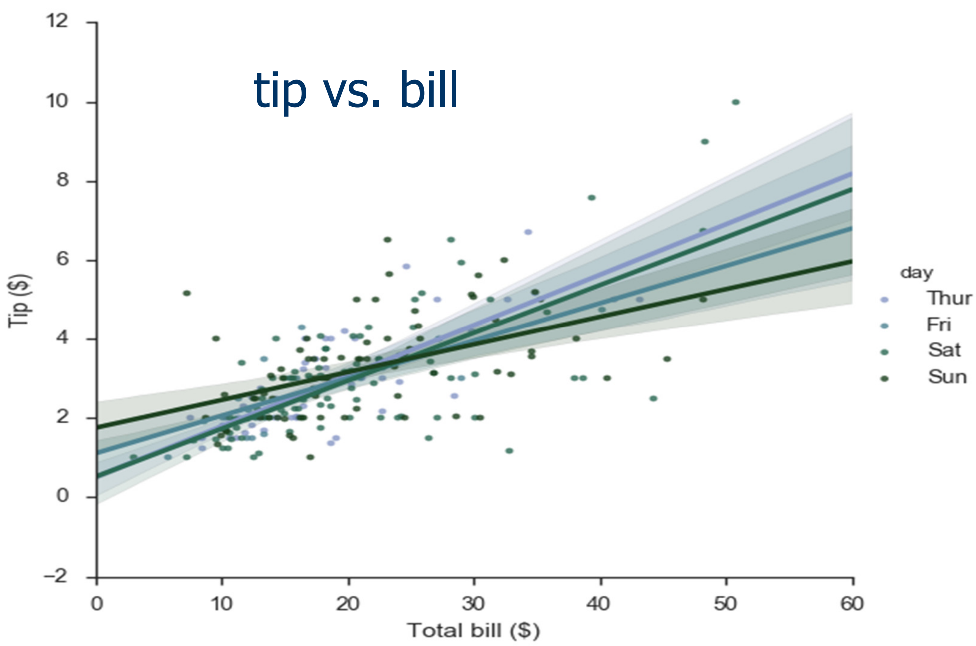
Término que engloba técnicas para el modelado y análisis de datos numéricos con el objetivo de predecir el valor de una variable a partir del valor de otras:

- **Variable dependiente,**
  variable de respuesta o medida:
  La variable cuyo valor deseamos predecir.

- **Variables independientes,**
  variables explicativas o factores:
  Variables cuyos valores usamos para predecir.

Análisis de regresión

y=f(x)



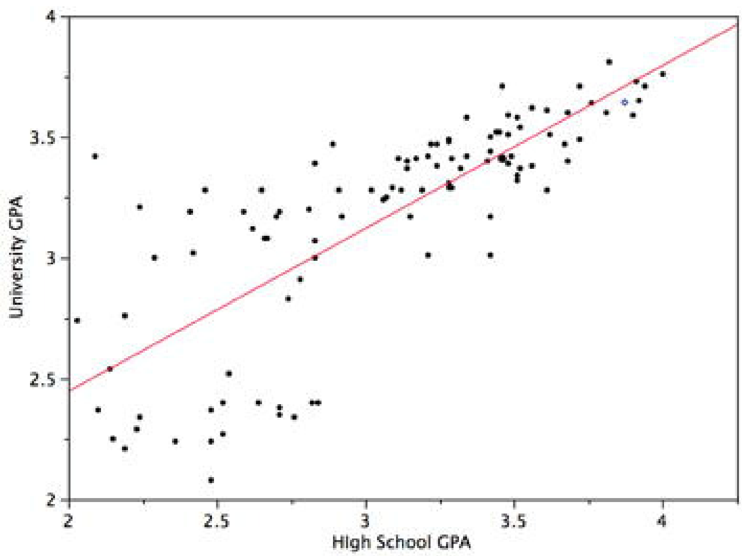
Técnicas de regresión

Las técnicas de regresión modelan la relación entre una o más variables independiente (predictores) y una variable dependiente (variable de respuesta).

Métodos de regresión

- Regresión lineal
- Regresión no lineal
- Regresión local
  - LOESS
  - Árboles de regresión (p.ej. CART)

Regresión lineal



Regresión lineal

Regresión lineal simple

Una única variable independiente:

$$y = w_0 + w_1 x$$

donde w_0 (desplazamiento) y w_1 (pendiente) son los coeficientes de regresión.

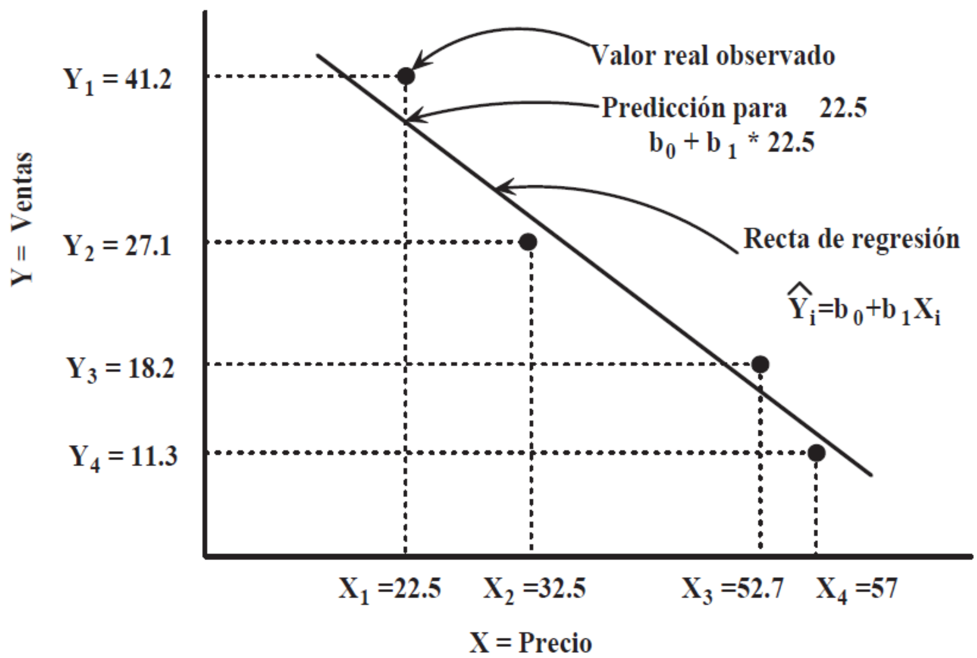
- **Método de los mínimos cuadrados**
  (estima la línea recta que mejor se ajusta a los datos):

$$w_0 = \bar{y} - w_1\bar{x} \qquad w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

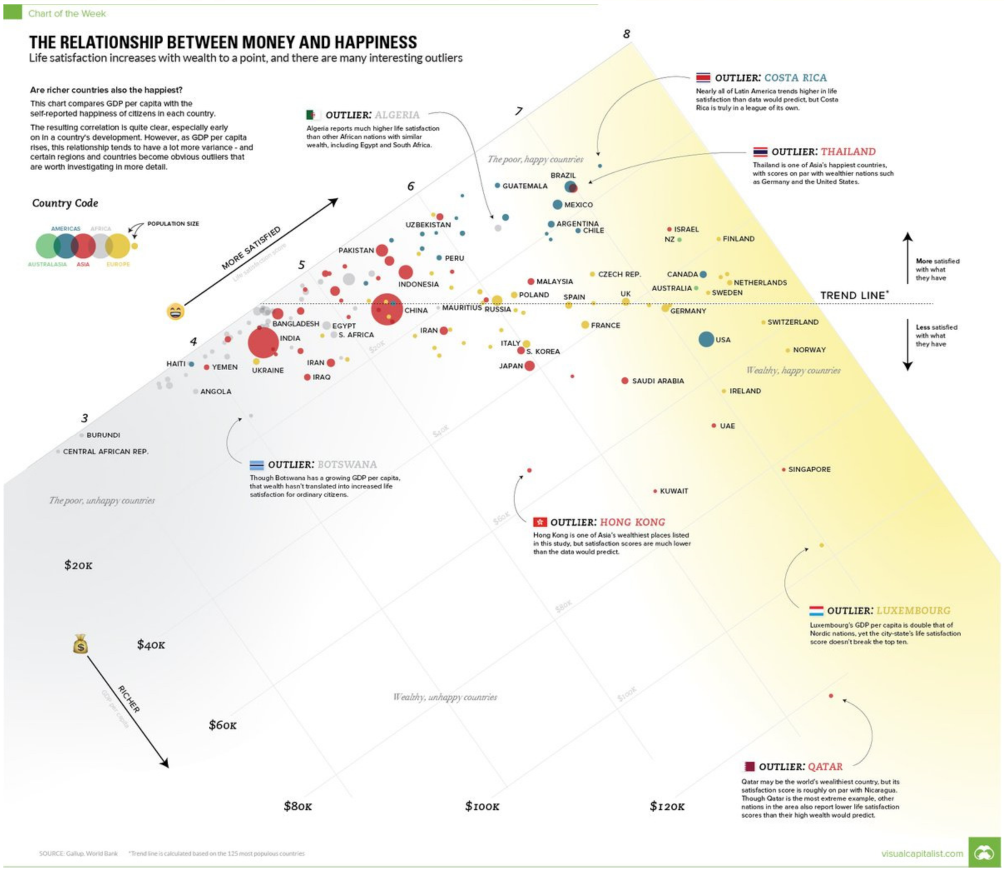
Regresión lineal

Regresión lineal simple

$$\hat{Y}_i = b_0 + b_1 X_i$$

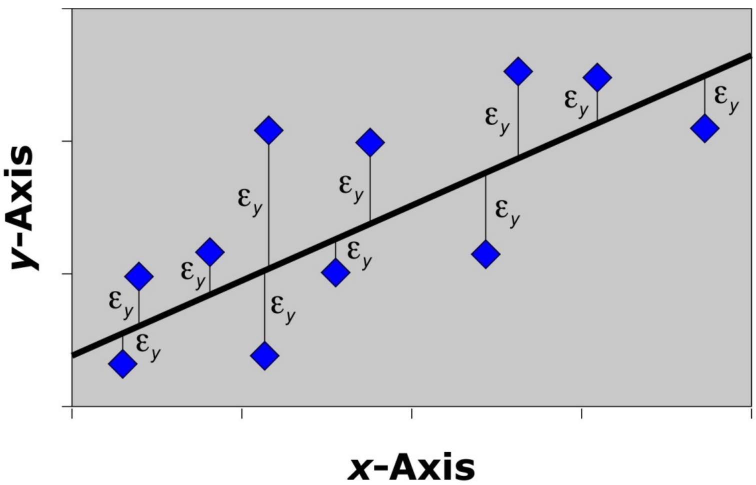


Regresión lineal



Regresión lineal

Regresión lineal simple

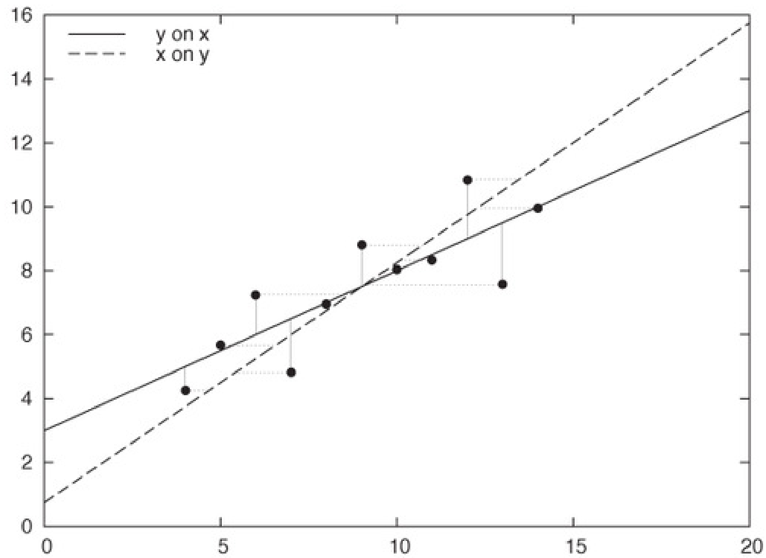


El método de los mínimos cuadrados minimiza la suma de los cuadrados de los residuos $\varepsilon_i = \hat{y}_i - y_i$ (las diferencias entre las predicciones y los valores observados). Formalmente:

$$\min F = \min \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Regresión lineal

Regresión lineal simple



¡OJO! Al utilizar regresión lineal, la recta y=f(x) que se obtiene es distinta a la que obtenemos si x=f(y).

Regresión lineal

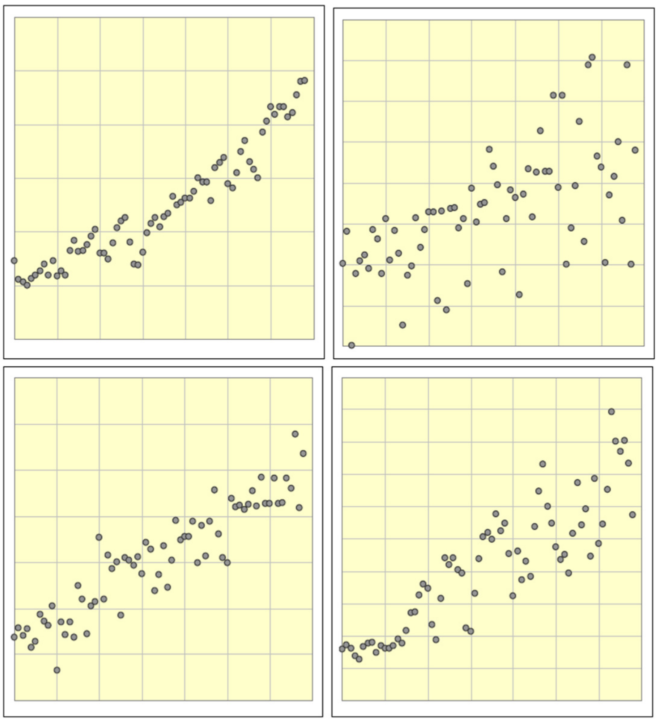
Suposiciones

Condiciones necesarias para aplicar regresión lineal:

- Obviamente, la muestra ha de ser aleatoria.
- El tipo de dependencia descrita ha de ser lineal.
- Fijado un valor de la(s) variable(s) independiente(s), la variable dependiente se distribuye según una distribución normal.
- Los errores han de tener la misma varianza (nube de puntos homogénea).

Regresión lineal

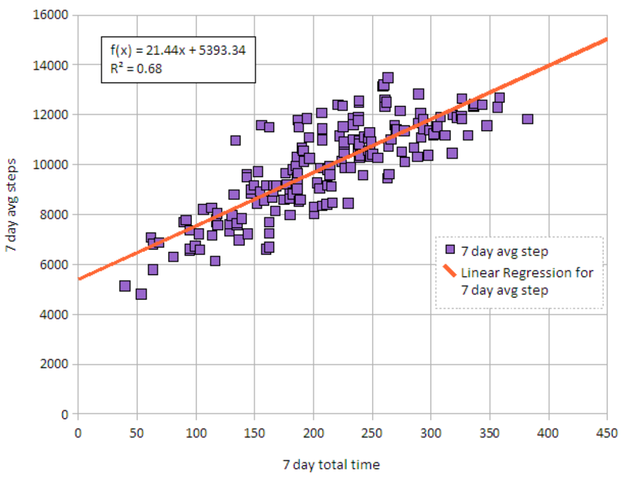
¿Qué suposiciones no se cumplen?



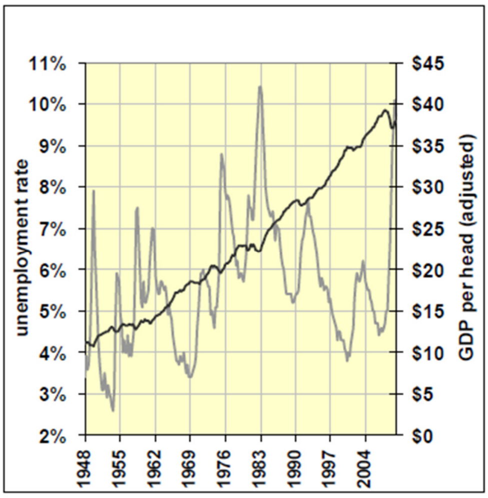
Regresión lineal

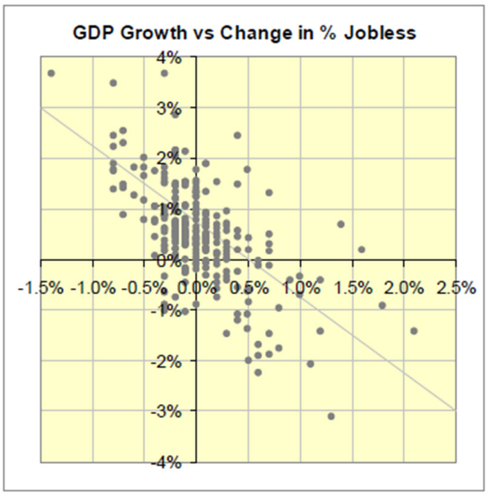
Regresión lineal simple

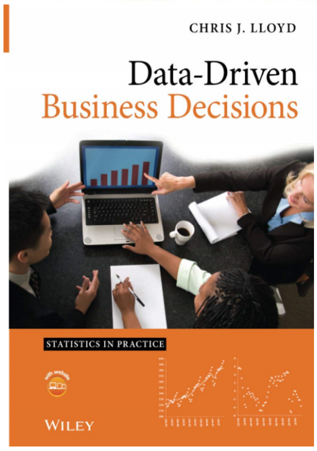
1. Mediante un diagrama de dispersión, comprobamos visualmente si existe una relación lineal entre las variables X (predictor) e Y (respuesta):



Regresión lineal simple







Ley de Okun: PIB per cápita y tasa de desempleo

GDP change = -1.372 x (unemployment rate change) + 0.543 %

Regresión lineal simple

2. Cuantificamos la relación construyendo la recta que resume la dependencia y damos una medida de cómo se ajusta la recta a los datos (correlación):

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \in [-1,1]$$

$$w_1 = r\frac{\sigma_y}{\sigma_x} \qquad w_0 = \bar{y} - w_1\bar{x}$$

Regresión lineal

Coeficiente de correlación lineal

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \in [-1,1]$$

r=+1 Dependencia lineal total en sentido positivo (cuanto mayor es X, mayor es Y).

r=-1 Dependencia lineal total en sentido negativo (cuanto mayor es X, menor es Y).

Regresión lineal

Coeficiente de correlación lineal

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \in [-1,1]$$

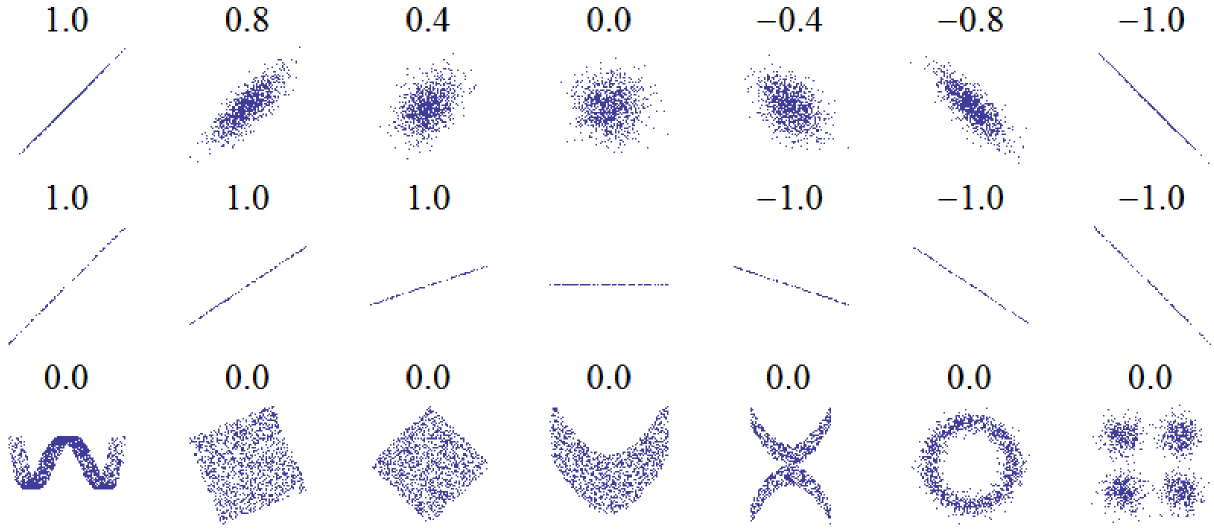
r > 0 Existe una dependencia positiva.
Cuanto más se acerque a 1, mayor es ésta.

r < 0 Existe una dependencia negativa.
Cuanto más se acerque a -1, mayor será.

r = 0 No podemos afirmar nada.

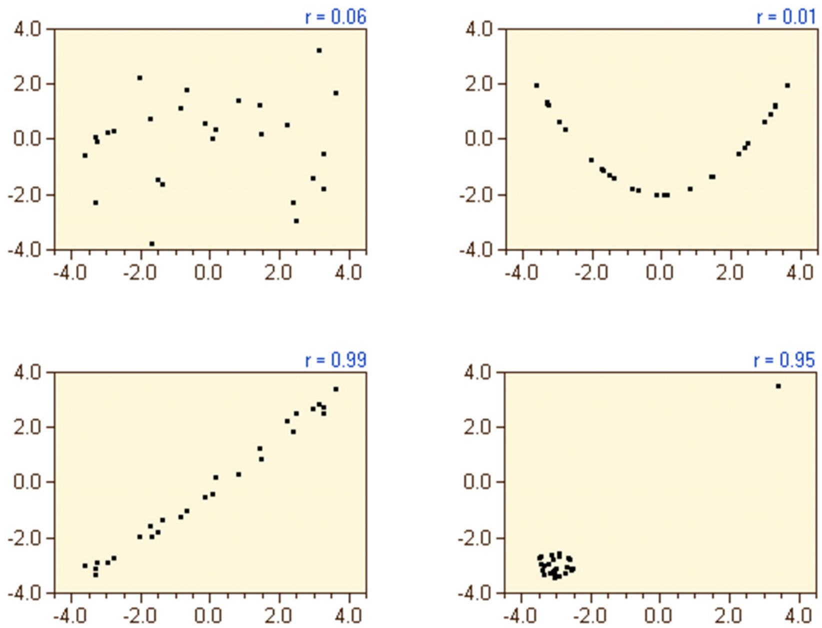
Regresión lineal

Coeficiente de correlación lineal



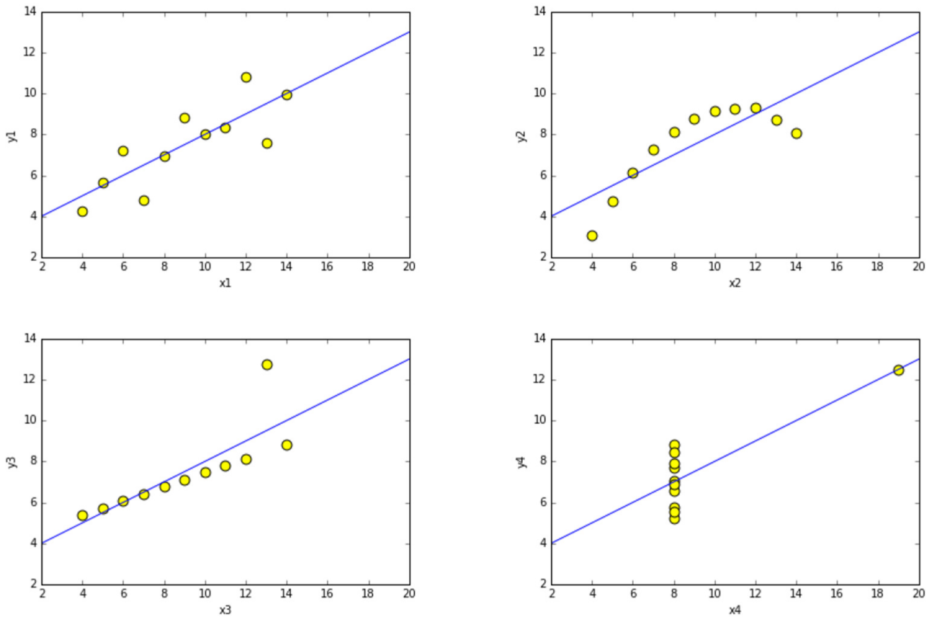
Regresión lineal

Coeficiente de correlación lineal



Regresión lineal

Coeficiente de correlación lineal



El cuarteto de Anscombe

(4 conjuntos de datos con el mismo coeficiente de correlación)

Regresión lineal

Coeficiente de correlación lineal

Ventaja de r

- No depende de las unidades usadas en la medición.

Limitaciones de r

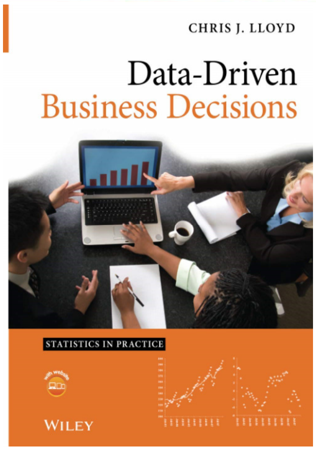
- Sólo mide dependencia lineal entre las variables.

¡OJO! La correlación no implica causalidad…

Regresión lineal

Coeficiente de correlación lineal

	YHOO	AOL	CSCO	MSFT	INTC	MRK	BAC	LUV	Viab	CT
YHOO		**0.47**	0.39	0.34	0.28	0.20	0.22	0.29	0.14	0.06
AOL	**0.47**		0.44	0.41	0.31	0.31	0.29	0.33	0.20	0.09
CSCO	0.39	0.44		**0.59**	**0.59**	0.41	0.36	0.36	0.21	0.16
MSFT	0.34	0.41	**0.59**		**0.62**	0.43	0.35	0.34	0.23	0.10
INTC	0.28	0.31	**0.59**	**0.62**		0.37	0.37	0.30	0.21	0.10
MRK	0.20	0.31	0.41	0.43	0.37		0.39	0.34	0.24	0.08
BAC	0.22	0.29	0.36	0.35	0.37	0.39		0.34	0.25	0.17
LUV	0.29	0.33	0.36	0.34	0.30	0.34	0.34		0.23	0.07
Viab	0.14	0.20	0.21	0.23	0.21	0.24	0.25	0.23		0.04
CT	0.06	0.09	0.16	0.10	0.10	0.08	0.17	0.07	0.04	

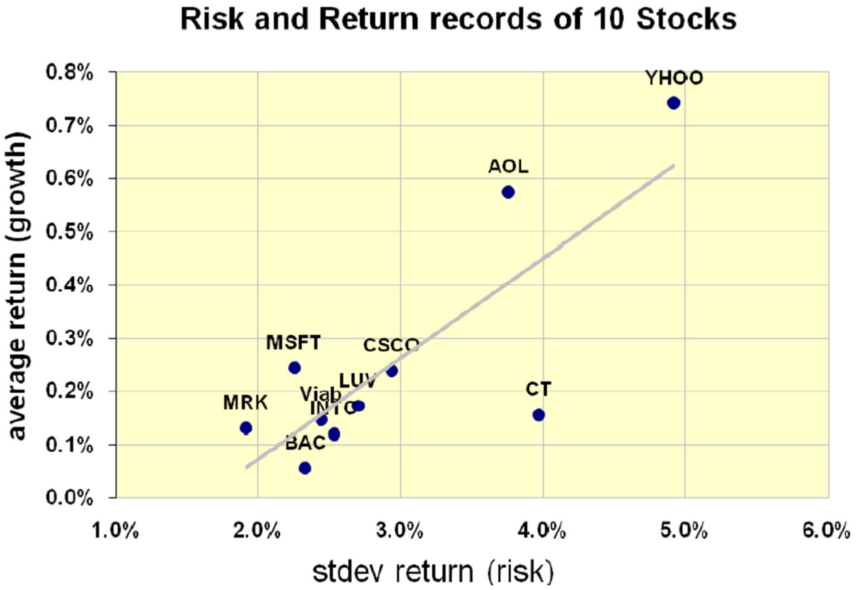


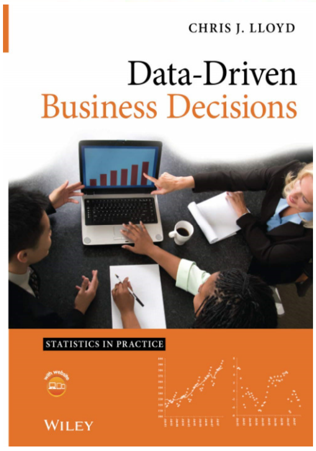
Correlación entre las cotizaciones de 10 empresas

- Mayor entre las empresas del mismo sector.
- Casi siempre positiva.

Regresión lineal

Coeficiente de correlación lineal

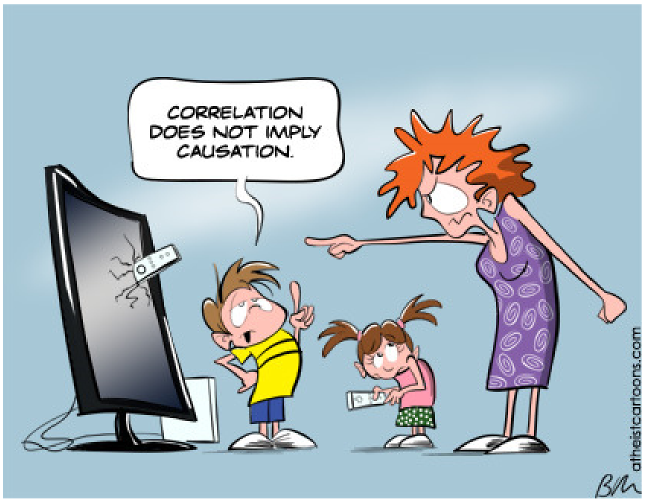




	YHOO	AOL	MSFT	INTC	MRK	BAC	CT
Mean (return)	0.74%	0.57%	0.24%	0.12%	0.13%	0.06%	0.16%
Stdev (risk)	4.91%	3.76%	2.25%	2.53%	1.91%	2.32%	3.97%

Regresión lineal

Coeficiente de correlación lineal



"Correlation is not causation but it sure is a hint."

-- Edward Tufte

Regresión lineal

Coeficiente de correlación lineal

Interpretación de su significado

Una forma habitual de interpretar una correlación r es interpretar r^2 como el "porcentaje de la variabilidad de una variable explicado por la otra variable"

p.ej. $r=0.9$ $r^2=0.81$
"81% de la variación de Y explicada por X"
$r=0.7$ $r^2=0.49$
"49% de la variación de Y explicada por X"

Regresión lineal

Coeficiente de correlación lineal

Error estándar de la correlación

$$SE(r) = \frac{\sqrt{1-r^2}}{\sqrt{n-2}}$$

Estadístico t: ¿es nula la correlación?

$$t = \frac{estimación\ - referencia}{SE(estimación)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Heurística: La correlación sólo es significativa si su magnitud es superior a 2/sqrt(n),
p.ej. con 100 datos (n=100) debería ser |r|>0.2

Regresión lineal

Error típico de la regresión

RMSE [Root Mean Square Error]

$$RMSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Bajo las suposiciones del modelo de regresión (linealidad, independencia, normalidad y variabilidad constante del error), RMSE nos da una estimación de la desviación estándar de la regresión (distancia típica de los puntos a la línea de regresión)…

Regresión lineal

Error típico de la regresión

RMSE [Root Mean Square Error]

Desviación estándar

$$\hat{\sigma} = RMSE$$

Error estándar de la pendiente de la recta

$$\hat{m} = r_{xy}\frac{\sigma_y}{\sigma_x} \qquad \mathrm{SE}(\hat{m}) = \frac{\hat{\sigma}}{\sigma_x\sqrt{n-1}}$$

NOTA: La raíz cuadrada de n aparece en el denominador:
para doblar la precisión de la estimación,
hemos de multiplicar por 4 el tamaño de la muestra.

Regresión lineal

Error típico de la regresión

Ley de Okun

GDP change = -1.372 x (unemployment rate change) + 0.543 %

RMSE = 0.796%

Error típico en el que incurriríamos si predecimos el cambio en el PIB del trimestre actual a partir del cambio en la tasa de desempleo.

SE(m) = 0.114

-1.372 ± 2 SE nos da un intervalo de confianza del 95%
para la pendiente real de recta (distribución normal).

Regresión lineal

Error típico de la regresión

Salida de una herramienta estadística

GDP change = -1.372 x (unemployment rate change) + 0.543 %

Summary measures				**Predicted Value:**		0.543%	
Adj RSquare		37.0%		Pred Error		0.799%	
Model Error		0.796%		Trend Error		0.071%	
p-value		0.0000					
Regression coefficients							
	Coefficient	Std Err	t-value	p-value	Lower	Upper	Partial
Constant	0.543%	0.051%	10.7	0.000	0.444%	0.643%	
UE Rate	-1.372	0.114	-12.1	0.000	-1.596	-1.148	-0.610

Regresión lineal

Preguntas a las que podemos responder

- Estimar el cambio en la salida Y si la entrada x cambia en una cantidad determinada Δx
- Estimar la salida media para todos los individuos que tengan un valor x.
- Predecir la salida real Y para un individuo particular x.

Podemos dar nuestra mejor estimación
y también evaluar la precisión de esa estimación.

Regresión lineal

Preguntas a las que podemos responder

1. Cambio en la salida Y si la entrada x cambia en Δx

$$\Delta Y = \widehat{m}\, \Delta x \qquad SE(\Delta Y) = SE(\widehat{m})\, \Delta x$$

2. Media de la tendencia

$$\mu(Y|x) = \widehat{m}x + \widehat{b} \qquad SE(\mu(Y|x)) = \hat{\sigma}\sqrt{\frac{1+z_x^2}{n}}$$

$$z_x = \frac{x-\overline{x}}{\sigma_x}$$

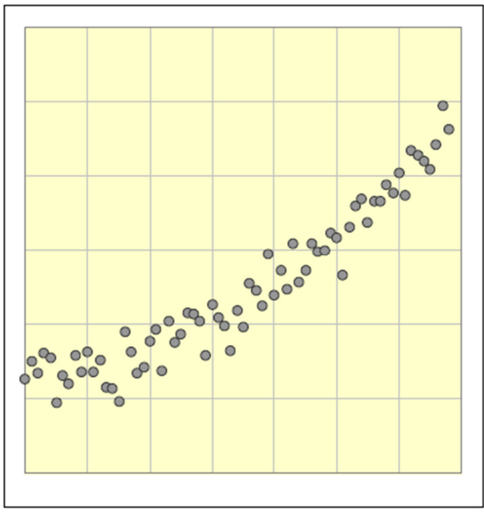
3. Predicción

$$\widehat{Y}\,(x) = \widehat{m}x + \widehat{b} \qquad SPE(\widehat{Y}) = \hat{\sigma}\sqrt{1+\frac{1+z_x^2}{n}}$$

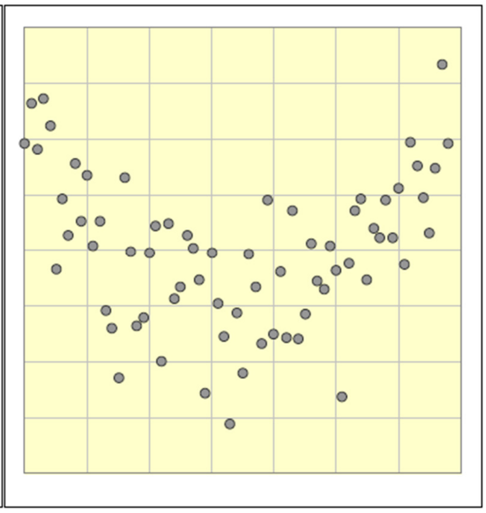
Regresión lineal

Análisis de residuos

Errores de predicción:
Diferencias entre el valor real y el valor ajustado



Datos originales

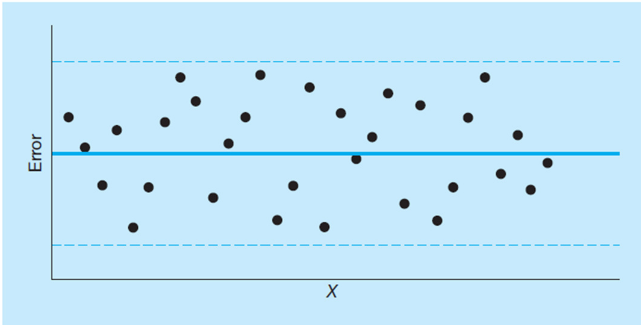


Residuos :-(

Regresión lineal

Análisis de residuos

FIGURE 4.4A
Pattern of Errors Indicating Randomness



Homocedasticidad

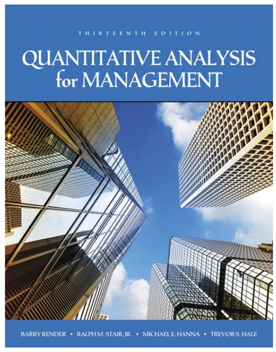
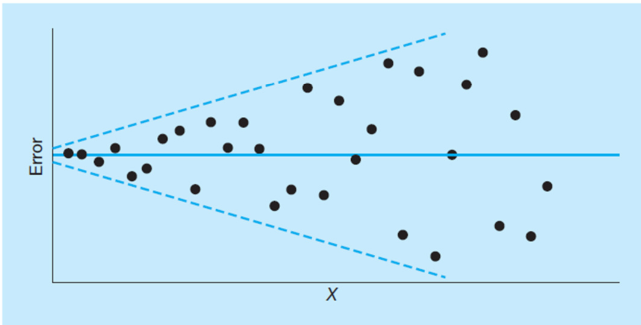


FIGURE 4.4B
Nonconstant Error Variance



Heterocedasticidad

Regresión lineal múltiple

Regresión lineal múltiple

Varias variables independientes:

$$y = w_0 + w_1\, x_1 + w_2\, x_2 + \ldots + w_p\, x_p$$

- Resoluble por métodos numéricos de optimización.
- Muchas funciones no lineales pueden transformarse en una expresión lineal.

p.ej. Un modelo de regresión polinomial
$y = w_0 + w_1\, x + w_2\, x^2 + w_3\, x^3$
puede transformarse en un modelo lineal definiendo las variables $x_2 = x^2$, $x_3 = x^3$:
$y = w_0 + w_1\, x + w_2\, x_2 + w_3\, x_3$

Regresión lineal múltiple

Suposiciones

$y = w_0 + \Sigma\, w_i\, x_i + \varepsilon$

- Linealidad
- Independencia (de los errores ε)
- Varianza (homocedasticidad): ε tiene la misma desviación σ, independientemente de los valores x_i
- Normalidad (de los errores ε, que rara vez se salen del intervalo -3σ a $+3\sigma$)

Regresión lineal múltiple

Ajuste del modelo de regresión

$y = w_0 + \Sigma\, w_i\, x_i + \varepsilon$

$$SSE = \sum_{j=1}^{n} (\hat{y}_j - y_j)^2 \qquad \hat{\sigma} = \sqrt{\frac{SSE}{n-p-1}}$$

$$\mathrm{SE}(\widehat{w_i}) = \frac{\hat{\sigma}}{\sigma_i \sqrt{n-1}}$$

Regresión lineal múltiple

Valores t y valores p [t-values & p-values]

El efecto de una variable x_i viene dado por su peso w_i

- Valor t (efecto observado en unidades estadísticas):

$$t = \frac{estimación - referencia}{SE(estimación)} = \frac{\widehat{w_i} - 0}{SE(\widehat{w_i})}$$

- Valor p (probabilidad de encontrarse un efecto tan grande como el observado por mero azar):

$$p = 2\,(1 - Normal(|t|))$$

Cuanto más grande t, más pequeño p.

Regresión lineal múltiple

Bondad del modelo de regresión

- $\hat{\sigma}$ (tamaño medio de los residuos):

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-p-1}} \qquad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **r^2** (coeficiente de determinación), más alto cuanto más se acerquen las predicciones a los valores reales.

  PROBLEMA: Si añadimos nuevas variables a nuestro modelo de regresión, el coeficiente de determinación r^2 aumenta.

  SOLUCIÓN: Se ajusta r^2 en función del número de casos y del número de variables. El valor r^2 ajustado siempre será algo menor que r^2.

$$r^2_{adjusted} = 1 - \frac{n-1}{n-p-1}(1 - r^2)$$

Regresión lineal múltiple

Coeficiente de determinación

$$r^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} = 1 - \frac{RSS}{TSS}$$

- TSS (varianza de los datos que la media no explica)

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

- RSS (error de predicción del modelo de regresión)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Regresión lineal múltiple

Coeficiente de determinación

$$r^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} = 1 - \frac{RSS}{TSS}$$

- TSS/n es la varianza de y
- RSS/TSS es la fracción de la varianza de y que el modelo de regresión es incapaz de explicar.
- Inversamente, 1 - RSS/TSS es la fracción de la varianza de y que el modelo es capaz de explicar.

Regresión lineal múltiple

Coeficiente de determinación

$$r^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} = 1 - \frac{RSS}{TSS}$$

- En el mejor caso, RSS=0
  y el coeficiente de determinación es uno.

- En el peor caso, RSS=TSS
  y el coeficiente de determinación es cero.

Regresión lineal múltiple

Coeficiente ajustado de determinación

$$r^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} = 1 - \frac{RSS/(n-1)}{TSS/(n-1)}$$

- Ajustamos 1-r^2 por un factor directamente proporcional al número de variables:

$$r^2_{adjusted} = 1 - \frac{RSS/df_{model}}{TSS/df_{mean\_model}}$$

$$= 1 - \frac{\frac{RSS}{n-1-p}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-p-1}\frac{RSS}{TSS}$$

$$= 1 - \frac{n-1}{n-p-1}(1-r^2)$$

Regresión lineal múltiple

Preguntas a las que podemos responder

1. Cambio en la salida Y si la entrada x_i cambia en Δx_i

$$\Delta Y = \widehat{w_i}\, \Delta x_i \qquad SE(\Delta Y) = SE(\widehat{w_i})\, \Delta x_i$$

2. Media de la tendencia

$$\mu(Y|x) = \widehat{w}x + \widehat{b} \qquad SE(\mu(Y|x)) = \hat{\sigma}\sqrt{\frac{p\bar{z}^2+1}{n}}$$

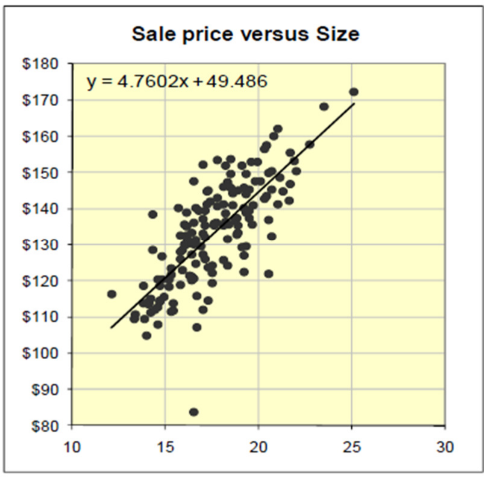
$$z_i = \frac{x_i - \overline{x_i}}{\sigma_{x_i}}$$

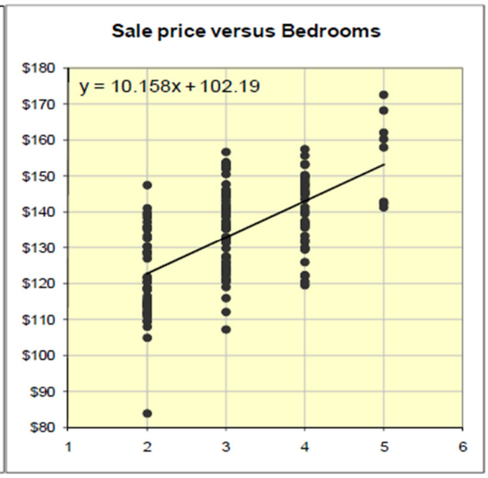
3. Predicción

$$\widehat{Y}\,(x) = \widehat{w}x + \widehat{b} \qquad SPE(\widehat{Y}) = \hat{\sigma}\sqrt{1+\frac{p\bar{z}^2+1}{n}}$$

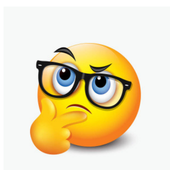
Regresión lineal múltiple

Ejemplo: Precio de un apartamento





Results of multiple regression for Price				
Regression coefficients				
	Coefficient	Std Err	t-value	p-value
Constant	49.095	6.747	7.3	0.000
Area	4.813	0.555	8.7	0.000
Beds	$\boxed{-0.178}$	1.491	-0.1	0.905



Multicolinealidad

Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i

Partimos del error observado

$$SSE = \sum_{j=1}^{n} (\hat{y}_j - y_j)^2 = \sum_{j=1}^{n} (w_o + \sum_{i=1}^{p} w_i x_{ji} - y_j)^2$$

Por conveniencia,
definimos x_0=1 y usamos notación vectorial:

$$SSE = \sum_{j=1}^{n} (\hat{y}_j - y_j)^2 = \sum_{j=1}^{n} (\sum_{i=0}^{p} w_i x_{ji} - y_j)^2 = \sum_{j=1}^{n} (w \cdot x_j - y_j)^2$$

Regresión lineal múltiple

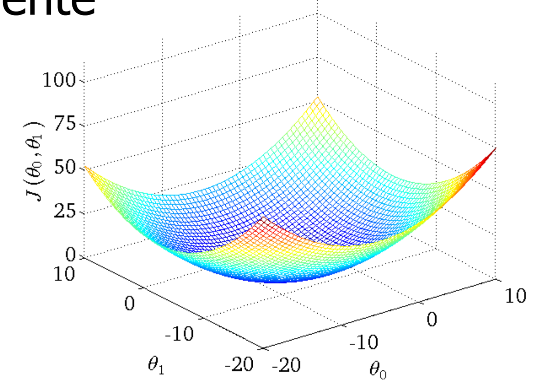
Cálculo de los coeficientes de regresión w_i

Minimizamos la función de coste

$$J(w) = \frac{1}{2n} SSE = \frac{1}{2n} \sum_{j=1}^{n} (\hat{y}_j - y_j)^2 = \frac{1}{2n} \sum_{j=1}^{n} (w \cdot x_j - y_j)^2$$

Usando el gradiente descendente

$$w_i := w_i - \alpha \frac{\partial}{\partial w_i} J(w)$$



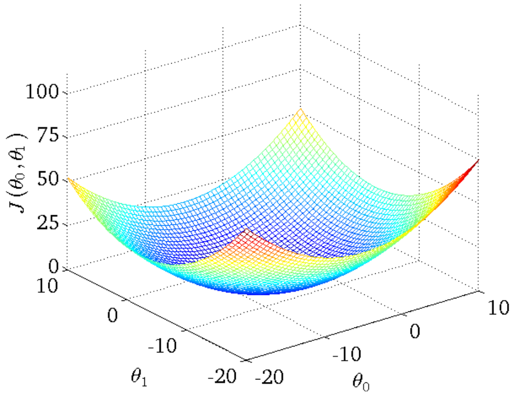
Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i

$$J(w) = \frac{1}{2n} \sum_{j=1}^{n} (w \cdot x_j - y_j)^2$$

$$\frac{\partial}{\partial w_i} J(w) = \frac{\partial}{\partial w_i} \frac{1}{2n} \sum_{j=1}^{n} \left(w \cdot x_j - y_j\right)^2 = \frac{1}{n} \sum_{j=1}^{n} (w \cdot x_j - y_j)\, x_{ji}$$

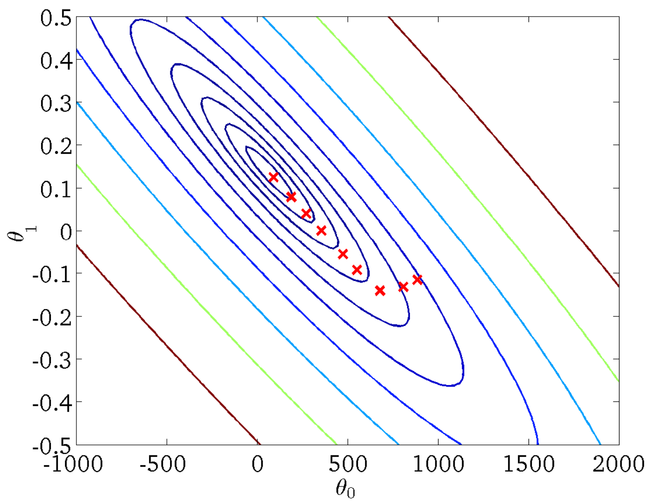
$$w_i \; := w_i - \alpha \frac{\partial}{\partial w_i} J(w)$$



Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i

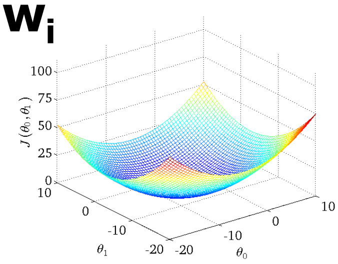
$$w_i \; := w_i - \alpha \frac{\partial}{\partial w_i} J(w)$$



[Andrew Ng]

Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i



$$w_i := w_i - \alpha \frac{\partial}{\partial w_i} J(w)$$

En la práctica:

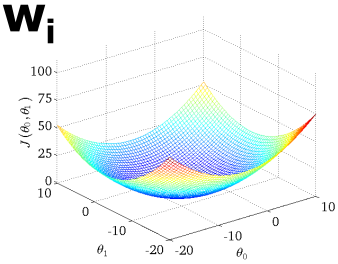
- Se normalizan las variables para agilizar la ejecución del algoritmo basado en el gradiente descendente.

$$x_j := \frac{x_j - \overline{x_j}}{\sigma_{x_j}} \quad \text{o bien} \quad x_j := \frac{x_j - \overline{x_j}}{max_{x_j} - min_{x_j}}$$

- La tasa de aprendizaje α ha de ser lo suficientemente pequeña para que la función de coste J se reduzca en cada iteración.

Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i



La ecuación normal

Resolución analítica de la regresión lineal

$$\frac{\partial}{\partial w_i} J(w) = 0$$

Matriz de diseño **X** de tamaño n x (p+1)

Vector de salidas **y** de tamaño n

Vector de predicciones **X·w** de tamaño n

Coste (ignorando constantes):

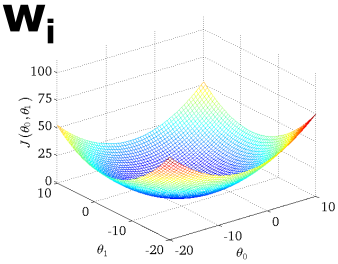
$$J(w) = (Xw - y)^T (Xw - y)$$

Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i

La ecuación normal

Resolución analítica de la regresión lineal



$$\frac{\partial}{\partial w} J(w) = \frac{\partial}{\partial w} (Xw - y)^T \, (Xw - y) = 0$$

$$\frac{\partial}{\partial w} J(w) = 2X^T \, (Xw - y) = 0$$

$$2X^T Xw - 2X^T y = 0$$

$$X^T Xw = X^T y$$

$$w = (X^T X)^{-1} \, X^T y$$

Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i

La ecuación normal

$$w = (X^T X)^{-1} \, X^T y$$

En Python:

```
import numpy as np

def linear_regression_normal_equation(X, y):
    X_transpose = np.transpose(X)
    X_transpose_X = np.dot(X_transpose, X)
    X_transpose_y = np.dot(X_transpose, y)
    try:
        return np.linalg.solve(X_transpose_X, X_transpose_y)
    except np.linalg.LinAlgError:
        return None
```

Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i

La ecuación normal

$$w = (X^T X)^{-1} X^T y$$

En la práctica,
la matriz $X^T X$ puede ser singular
(esto es, no ser invertible),
por lo que numéricamente se utiliza la pseudoinversa A^+
($A\,A^+\,A = A$), que siempre existe.

En Matlab:

```
w = pinv(X'*X)*X'*y
```


Regresión lineal múltiple

Cálculo de los coeficientes de regresión w_i

Gradiente descendente vs. Ecuación normal

$$w = (X^T X)^{-1} X^T y$$

Gradiente descendente	Ecuación normal
Hay que elegir el valor del hiperparámetro α	No hay que elegir el valor de ningún hiperparámetro
Necesita múltiples iteraciones	Es un método directo
Funciona bien incluso cuando n es grande	Requiere invertir una matriz de tamaño n x n, i.e. $O(n^3)$
O(knp)	**$O(n^3+n^2p)$**
Escalable	No escalable

Regresión lineal ++

Autocorrelación

Correlación de una serie temporal consigo misma (desplazada en el tiempo): sirve como medida de hasta qué punto los datos de una serie temporal pueden predecirse a partir de los valores previos.

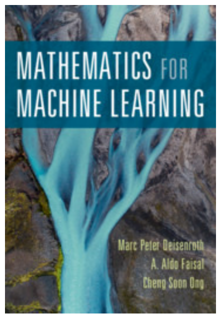
Correlaciones parciales

¿Hasta qué punto tiende Y a variar con X cuando las demás variables no cambian? Es algo que no puede verse directamente en la nube de puntos X-Y

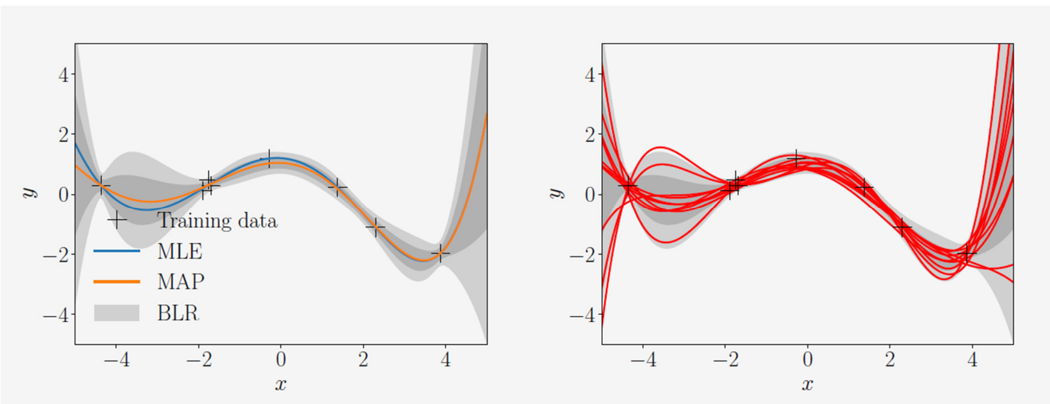
NOTA: Las correlaciones habituales
se denominan correlaciones marginales

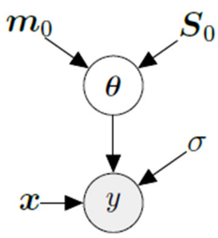
Regresión lineal ++

Regresión lineal bayesiana



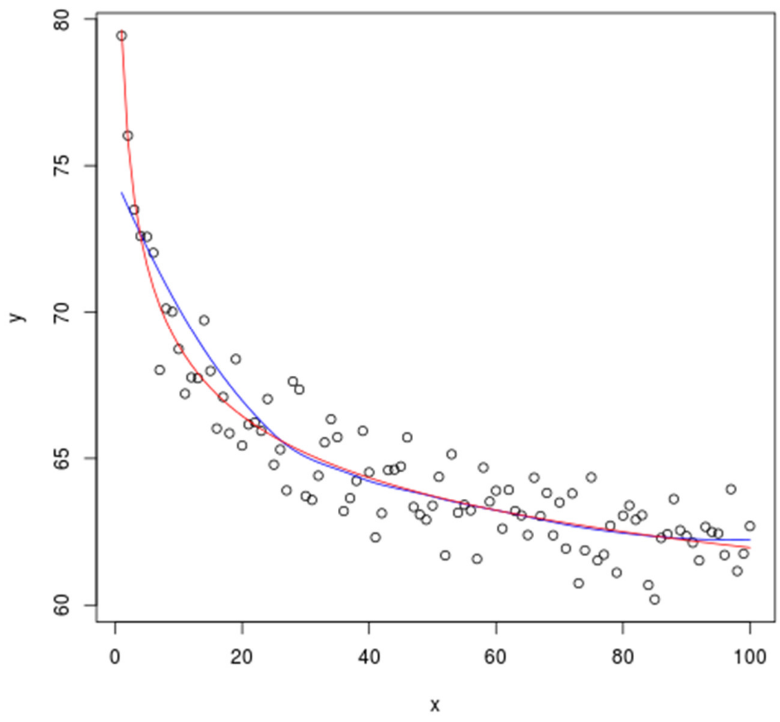
Nos permite modelar la incertidumbre de las predicciones realizadas…

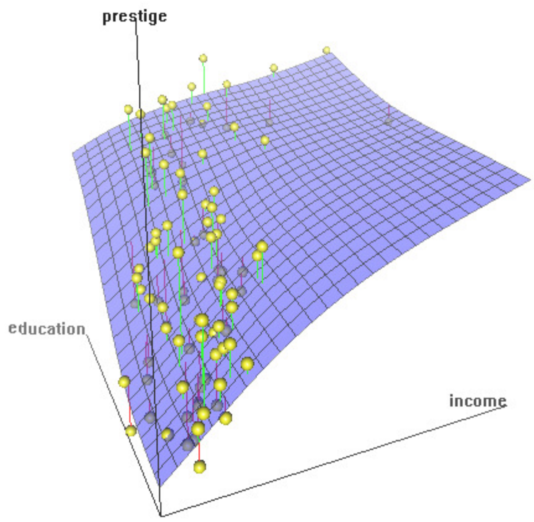




MLE: Maximum Likelihood Estimate
MAP: Maximum A Posteriori Estimate
BLR: Bayesian Linear Regression

Regresión no lineal





Regresión no lineal

Podemos añadir términos adicionales a la regresión:

- Variables binarias para indicar la presencia de una característica concreta [dummy variables],
  p.ej. rasgos binarios (0, 1)
  segmentación en grupos (dummy por grupo)

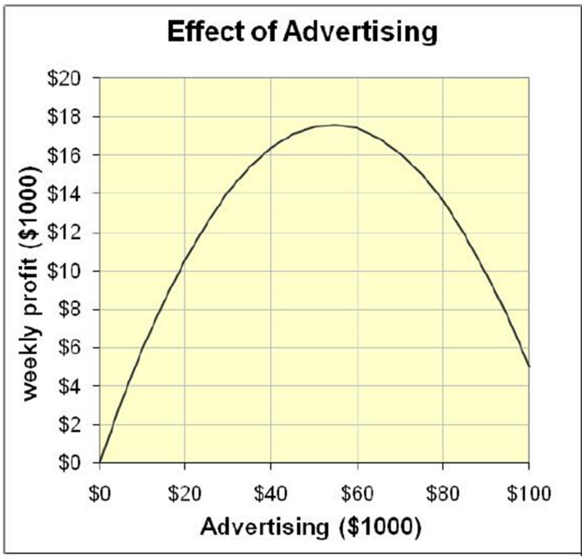
- Términos cuadráticos o logarítmicos para representar efectos curvilíneos, cuadráticos o proporcionales,
  p.ej. retorno de una inversión publicitaria
  depreciación de un activo
  efecto de los precios en la demanda

- Entradas que modelen interacciones entre variables diferentes [interaction variables], p.ej. X_1X_2

Regresión no lineal

EJEMPLO: Regresión cuadrática

$$Y = b + m_L x + m_Q x^2$$



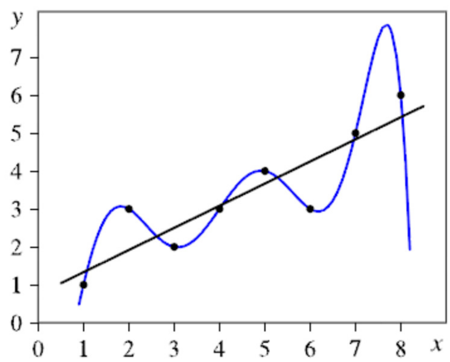
Óptimo: $x_{best} = -0.5 * m_L / m_Q$

Regresión no lineal

EJEMPLO: Regresión polinómica

$$Y = b + m_1 x + m_2 x^2 + m_3 x^3 + \ldots + m_p x^p$$

Es importante ajustar
el grado del polinomio p,
ya que dados n puntos
siempre se pueden encontrar
un polinomio de grado n-1
que pasa por todos ellos.



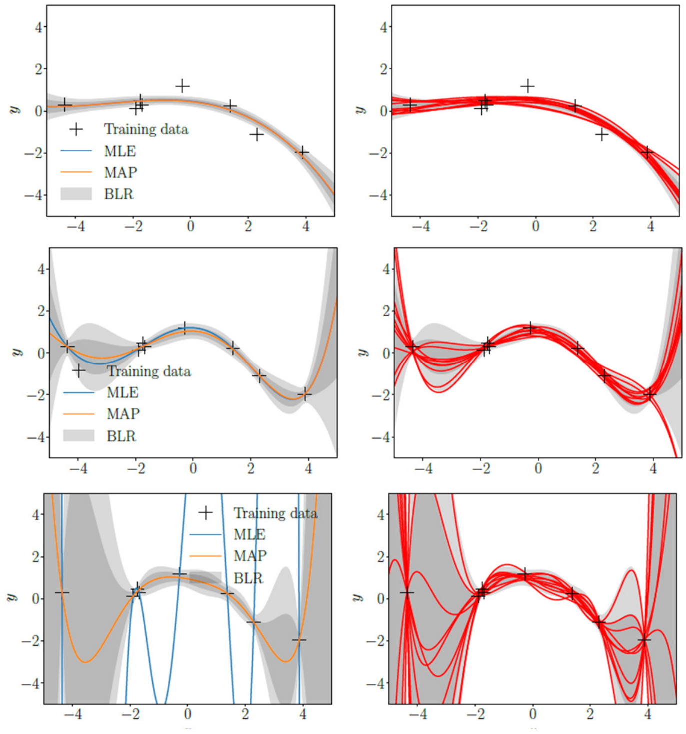
Regresión no lineal

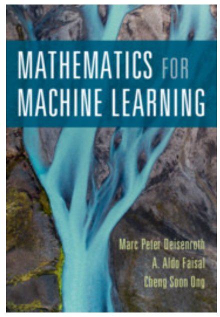
Aumento de la incertidumbre conforme aumenta el grado del polinomio (y del riesgo de sobreaprendizaje):

p=3

p=5

p=7





Regresión no lineal

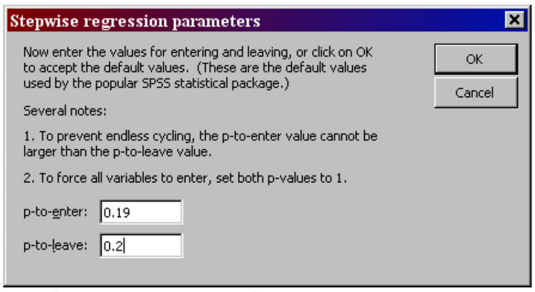
El número de términos que podríamos incluir es muy elevado (crece exponencialmente si admitimos combinaciones de variables), por lo que resulta útil emplear alguna técnica de selección de características:

- **Selección hacia adelante [forward selection]:**
  Se empieza con la variable que tenga mayor correlación con la salida Y. En cada paso, se añade la variable con menor valor P.

- **Eliminación hacia atrás [backward elimination]:**
  Se van eliminando las variables menos significativas mientras que su valor P sea mayor que un umbral (p.ej. P>0.2).

Regresión no lineal

EJEMPLO: Precio de un apartamento



Results of multiple regression for PRICE				
Summary measures				
Adj R-Square	83.0%			
Model Error	108.8			
Regression coefficients				
	Coefficient	Std Err	t-value	p-value
Constant	1147.0	148.3	7.7	0.000
SQFT	-677.7	166.6	-4.1	0.000
SQFT2	298.5	46.7	6.4	0.000
CARS	133.9	35.7	3.8	0.000
AGE	-26.0	10.0	-2.6	0.011
AGE2	0.7	0.4	1.6	0.111
FLOOR	9.8	4.2	2.3	0.023
BATHS_2	69.6	28.0	2.5	0.015
BATHS_3	93.0	60.3	1.5	0.126

Regresión no lineal

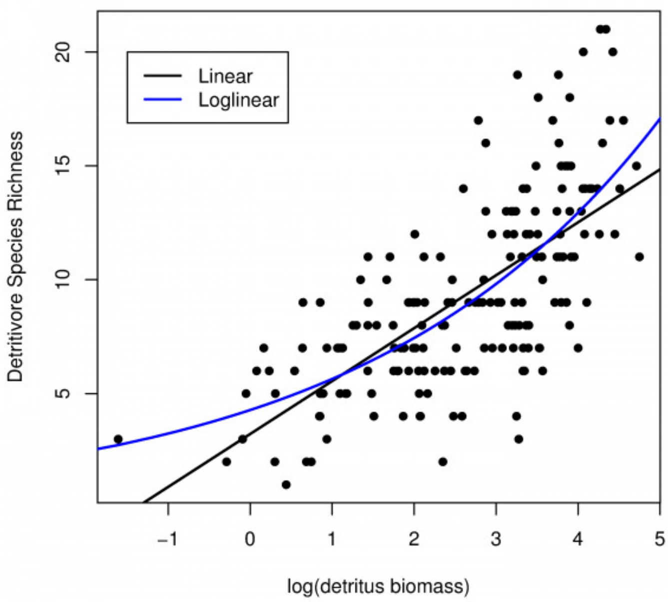
EJEMPLO: Precio de un apartamento

El precio por área aumenta con su altura (FLOOR*AREA)

Results of multiple regression for PRICE				
Summary measures				
Adj R-Square	95.3%			
Model Error	57315			
Regression coefficients				
	Coefficient	Std Err	t-value	p-value
Constant	1062921	81525	13.0	0.000
Time	1803	116	15.5	0.000
SQFT	-713	92.8	-7.7	0.000
SQFT2	0.300	0.0	11.5	0.000
CARS	172696	19408	8.9	0.000
AGE	-31503	5395	-5.8	0.000
AGE2	913	225	4.1	0.000
BEDS_3	24908	12808	1.9	0.055
BATHS_2	52477	15515	3.4	0.001
BATHS_3	88265	32425	2.7	0.008
FLOOR*AREA	2.89	1.29	2.2	0.027

Modelos log-lineales

y = ax^b → ln y = ln a + b ln x

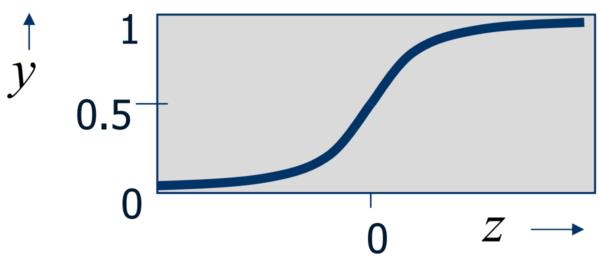


Regresión logística

Para resolver problemas de clasificación, preferiríamos obtener una salida binaria {0,1}

- Regresión lineal

$$\hat{y}_j = w \cdot x_j = \sum_{i=0}^{p} w_i x_{ji}$$



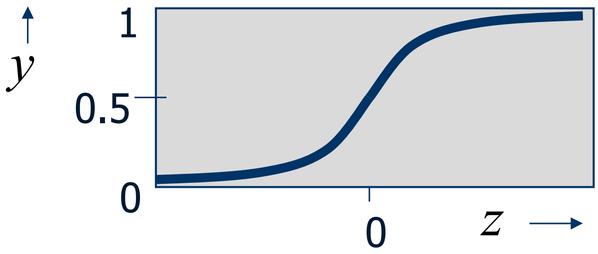
- Regresión logística

$$\hat{y}_j = f(w \cdot x_j) = f\left(\sum_{i=0}^{p} w_i x_{ji}\right) \qquad f(z) = \frac{1}{1+e^{-z}}$$

Regresión logística

Predicción a la hora de clasificar

- Clase positiva (y=1) $f(w \cdot x_j) \geq 0.5$
- Clase negativa (y=0) $f(w \cdot x_j) < 0.5$



La salida de la función logística también puede interpretarse como una probabilidad:

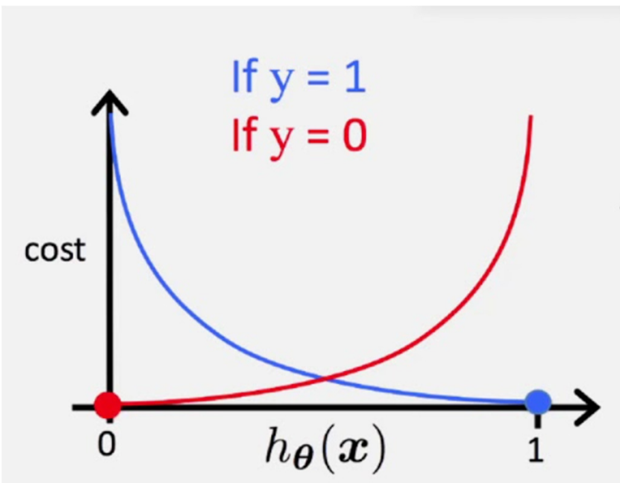
$$P(y = 1|x) = f(w \cdot x_j)$$

$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - f(w \cdot x_j)$$

Regresión logística

Función de coste asociada a la regresión logística

$$coste(f(w \cdot x_j), y) = \begin{cases} -\log\left(f(w \cdot x_j)\right) & si \quad y = 1 \\ -\log\left(1 - f(w \cdot x_j)\right) & si \quad y = 0 \end{cases}$$



Regresión logística

Función de coste asociada a la regresión logística

$$coste\left(f\left(w \cdot x_j\right), y\right) = \begin{cases} -\log\left(f\left(w \cdot x_j\right)\right) & si \quad y = 1 \\ -\log\left(1 - f\left(w \cdot x_j\right)\right) & si \quad y = 0 \end{cases}$$

$$J(w) = \frac{1}{n} \sum_{j=1}^{n} coste\left(f\left(w \cdot x_j\right), y\right)$$

Como y=0 o y=1, podemos simplificar:

$$coste\left(f\left(w \cdot x_j\right), y\right) = -y \log\left(f\left(w \cdot x_j\right)\right) - (1 - y)\log\left(1 - f\left(w \cdot x_j\right)\right)$$

Regresión logística

Para encontrar los parámetros del modelo minimizamos la función de coste:

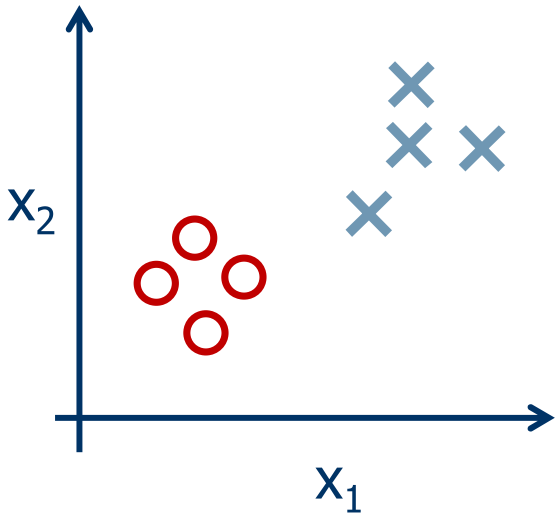
$$J(w) = \frac{1}{n} \sum_{j=1}^{n} \left[-y \log\left(f\left(w \cdot x_j\right)\right) - (1 - y)\log\left(1 - f\left(w \cdot x_j\right)\right)\right]$$

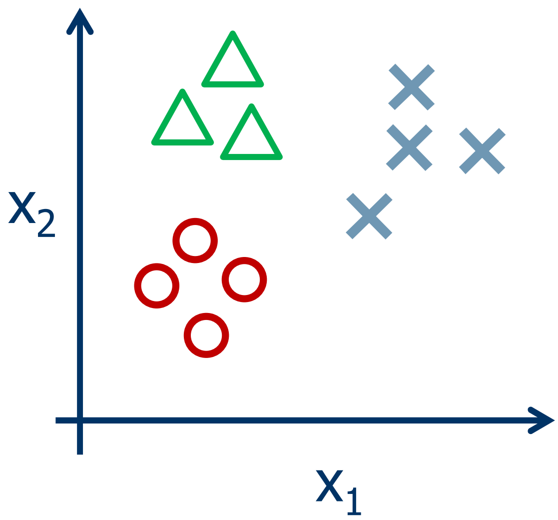
$$\frac{\partial}{\partial w_i} J(w) = \frac{1}{n} \sum_{j=1}^{n} \left[f\left(w \cdot x_j\right) - y_j\right] x_{ji}$$

Algoritmo idéntico a la regresión lineal.

Regresión logística

¿Y si tenemos más de dos clases?





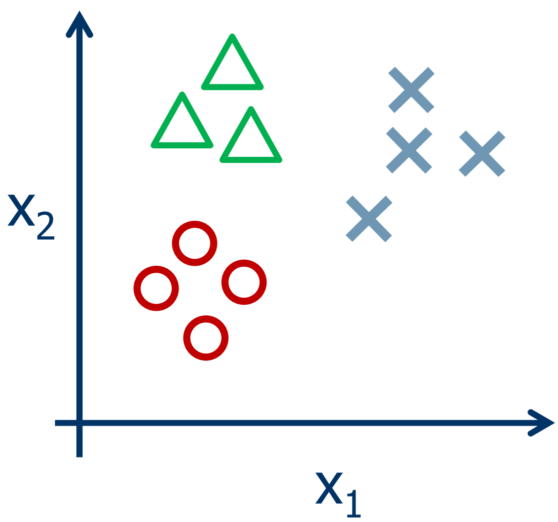
Clasificación binaria

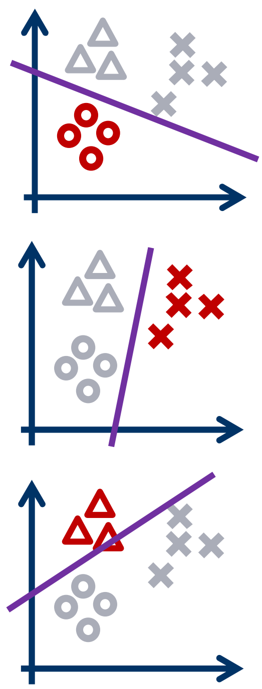
Clasificación multiclase

Regresión logística

¿Y si tenemos más de dos clases?

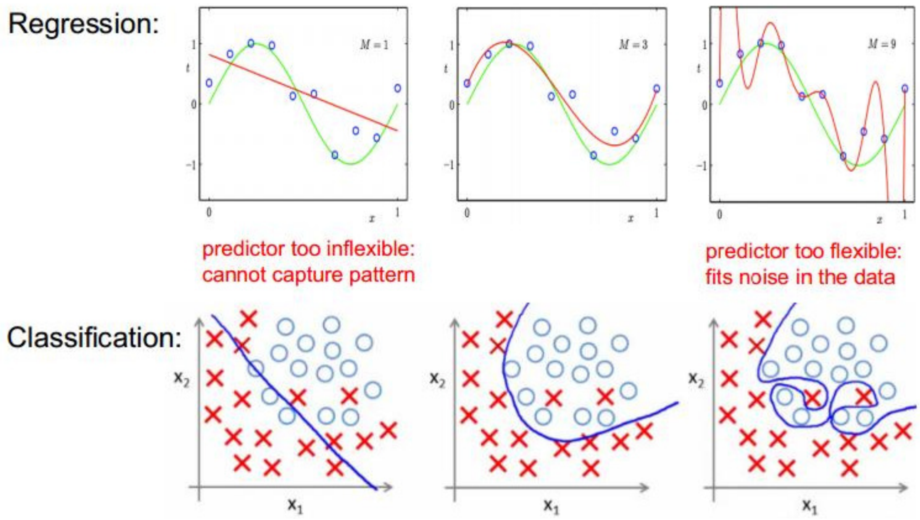
Creamos n clasificadores binarios [one-vs-all / one-vs-rest]





Regularización

El problema del sobreaprendizaje [overfitting]



El modelo puede ajustarse demasiado bien
al conjunto de datos de entrenamiento
pero ser incapaz de generalizar correctamente.

Regularización

¿Cómo podemos evitar el sobreaprendizaje?

- Reduciendo el número de características (selección de variables y extracción de características).
- Regularizando los parámetros del modelo (se mantienen todas las variables de entrada, pero se reduce la magnitud de los parámetros w_i).

Regularización

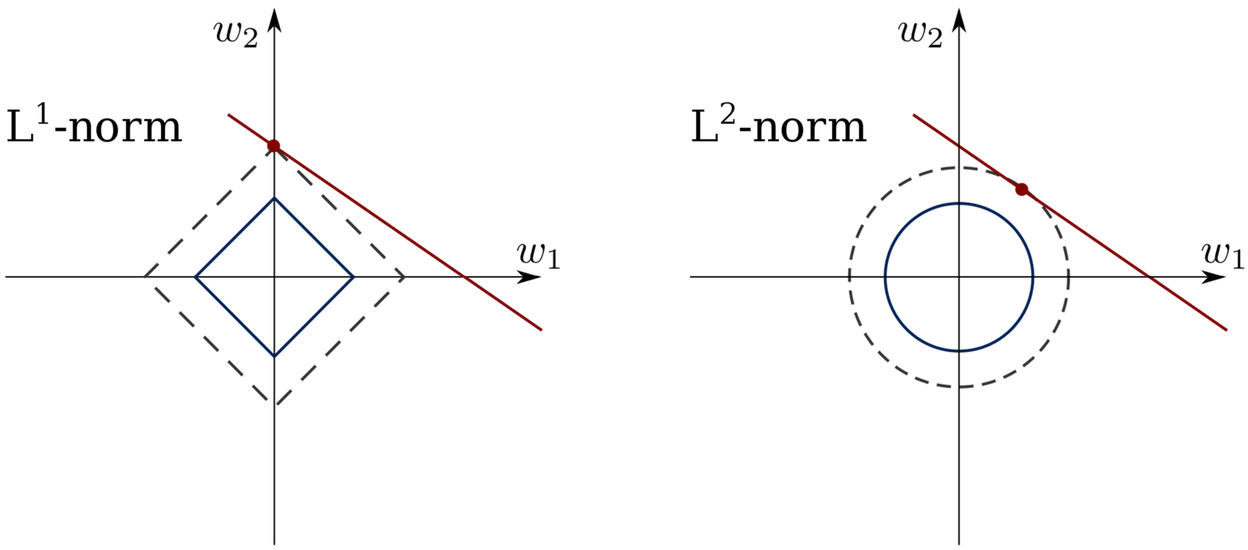
¿Cómo regularizamos los parámetros del modelo?

Introduciendo términos adicionales en la función de coste.

- **Regularización L2** [ridge/Tikhonov regression] $$J(w) = \frac{1}{n}\sum_{j=1}^{n} error_j + \lambda \sum_{i=1}^{p} w_i^2$$
- **Regularización L1** [lasso regression] $$J(w) = \frac{1}{n}\sum_{j=1}^{n} error_j + \lambda \sum_{i=1}^{p} |w_i|$$

Nuevo hiperparámetro: Factor de regularización λ

Regularización



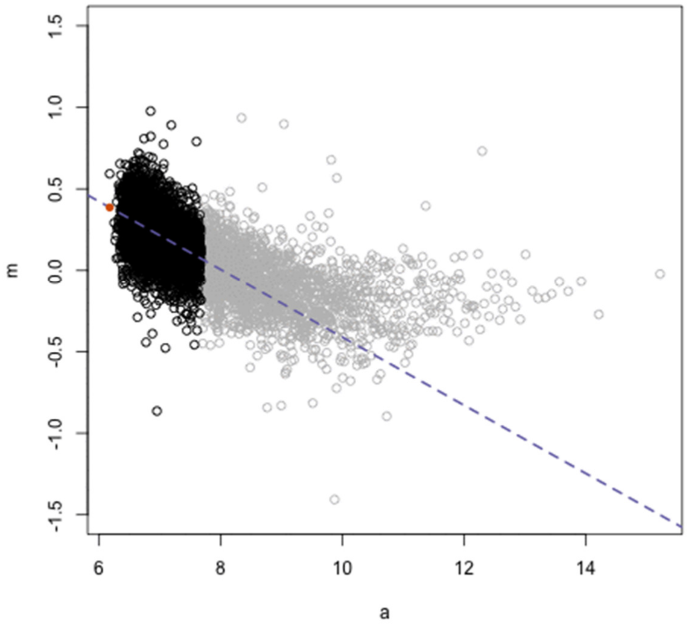
Se pueden combinar ambas:

Elastic net regularization (2005)

$$J(w) = \frac{1}{n}\sum_{j=1}^{n} error_j + \lambda_2 \sum_{i=1}^{p} w_i^2 + \lambda_1 \sum_{i=1}^{p} |w_i|$$

Técnicas de regresión local

LOESS [locally estimated scatterplot smoothing]



Técnicas de regresión local

LOESS [locally estimated scatterplot smoothing]

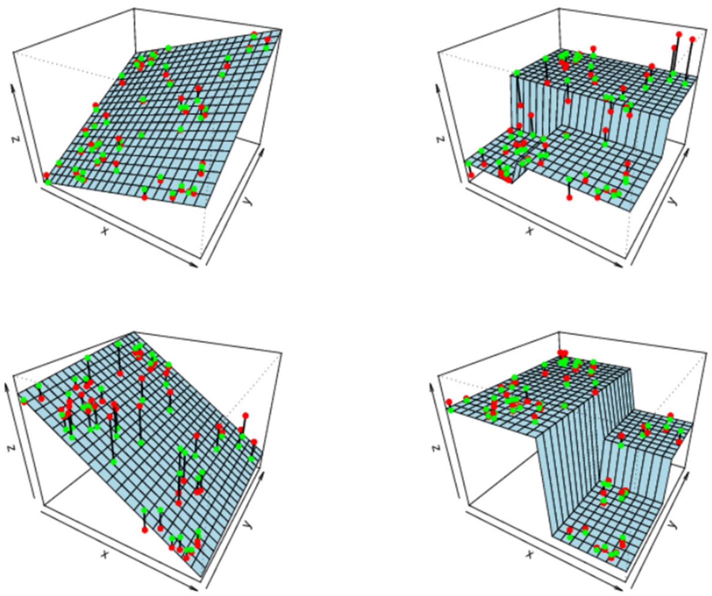
William S. Cleveland, 1979

Método de regresión no paramétrico que combina un modelo de regresión multiple con un metamodelo basado en los vecinos más cercanos (k-NN):

Combina la sencillez de la regresión lineal
con la flexibilidad de la regresión no lineal
ajustando modelos simples a subconjuntos de los datos.

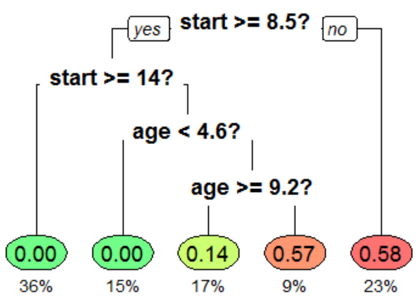
a.k.a. filtro Savitzky-Golay (1964)

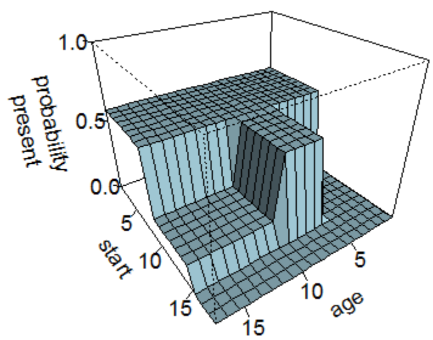
CART [Classification and Regression Trees]

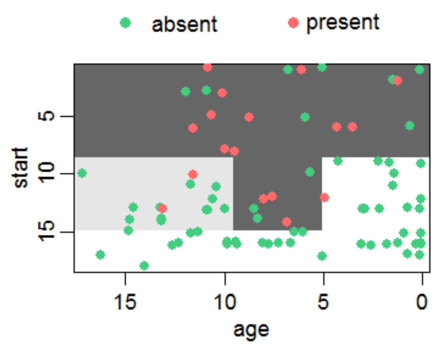


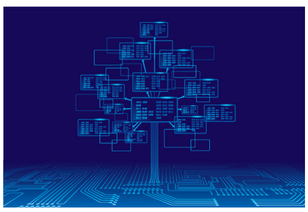
Regresión lineal CART

CART [Classification and Regression Trees]







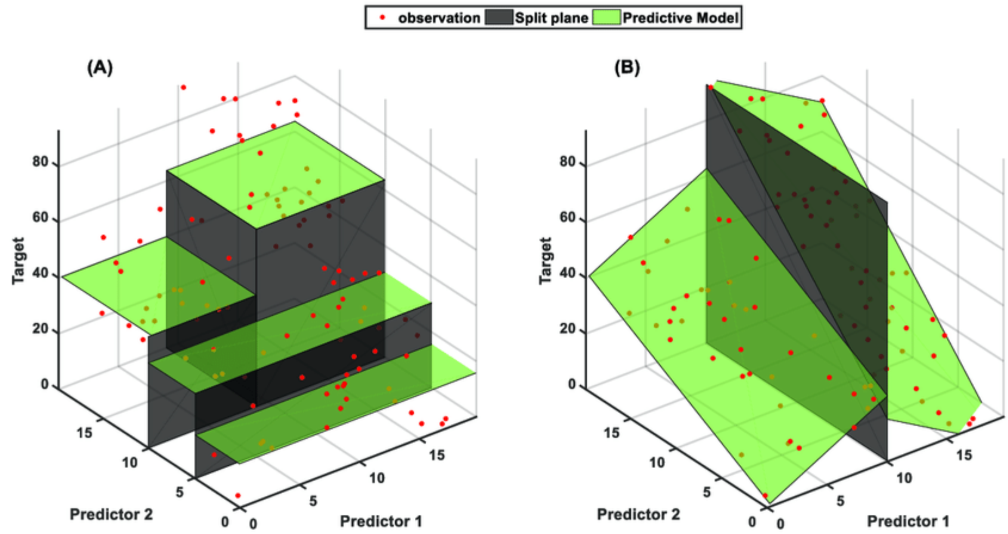


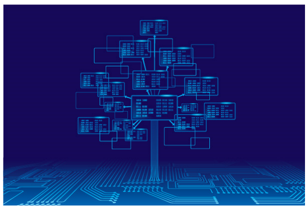
CART®

SPM® CART® Random Forests® MARS® TREENET®

Técnicas de regresión local

CART [Classification and Regression Trees]





Técnicas de regresión

Más técnicas de predicción…

Pedro Domingos
@pmddomingos

These days, instead of a linear regression you can use an LLM that will cost you a million times more and give you worse results.

Forecasting

http://en.wikipedia.org/wiki/Forecasting

Apéndice
Resultados de la regresión

Error

Diferencia entre el valor real y la predicción

$$error = y_i - \hat{y}_i$$

SST [Sum of Squares Total]

Variabilidad de y en torno a su media

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

SSE [Sum of Squared Error]

Variabilidad de y en torno a su línea de regresión

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$SSE < SST$

Apéndice
Resultados de la regresión

SSR [Sum of Squares due to Regression]

Variabilidad de y explicada por el modelo de regresión

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$SST = SSE + SSR$

R^2 [Determination Coefficient]

- SSR: Variabilidad explicada
- SSE: Variabilidad no explicada

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Interpretación:
Porcentaje de la variabilidad explicada por la regresión.

Resultados de la regresión

MSE [Mean Squared Error]

Varianza del error de regresión

$$s^2 = MSE = \frac{SSE}{n-p-1}$$

Error estándar de la estimación / Desviación estándar de la regresión

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p-1}}$$

Resultados de la regresión

Test de hipótesis para comprobar si realmente existe una relación lineal entre X e Y:

MSR [Mean Squared Regression]

$$MSR = \frac{SSR}{p}$$

Estadístico F

$$F = \frac{MSR}{MSE}$$

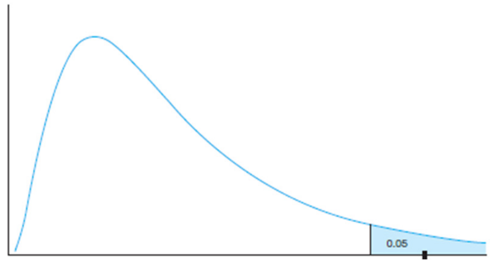
- Hipótesis nula: No existe relación lineal.
- Si el valor F es elevado, su valor p asociado será bajo y podremos rechazar la hipótesis nula (p.ej. $p<0.05$).

NOTA El test t evalúa cada coeficiente por separado, el test F evalúa el modelo en su conjunto.

Apéndice

Resultados de la regresión

Regression	
R Square	r^2
Adjusted R Square	r^2 ajustado
Standard Error	s
Observations	n



ANOVA	df	SS	MS	F	Significance
Regression	p	SSR	MSR = SSR/p	F	p-value
Residual	n-p-1	SSE	MSE = SSE/(n-p-1)		
Total	n-1	SST			

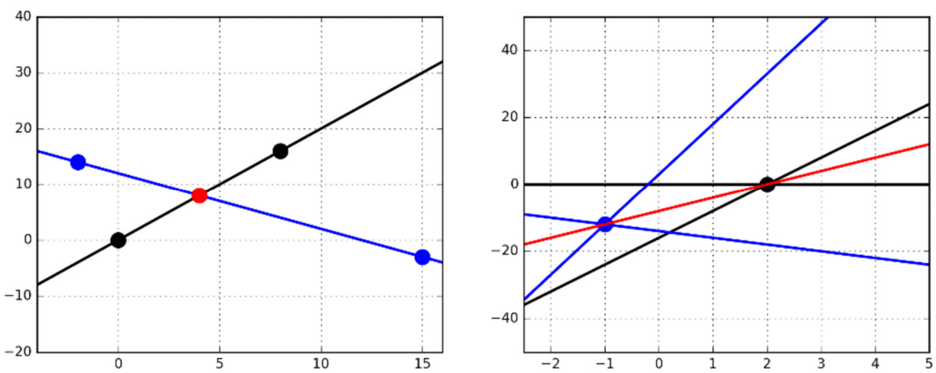
	Coefficient	Standard Error	t	p	Lower 95%	Upper 95%
Variable i	w_i	$SE(w_i)$	t	p	$\sim w_i-2SE(w_i)$	$\sim w_i+2SE(w_i)$
…						

Apéndice

Dualidad

Conexión entre regresión lineal
y resolución de sistemas de ecuaciones lineales

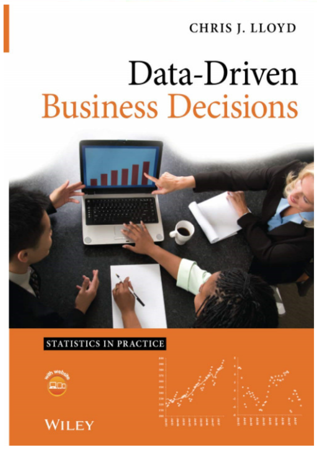
- Sistema lineal:
  Encontrar punto por el que pasan n líneas.
- Regresión:
  Dados n puntos, encontrar la línea que pasa por ellos.



(s,t) ↔ y = sx-t

Bibliografía

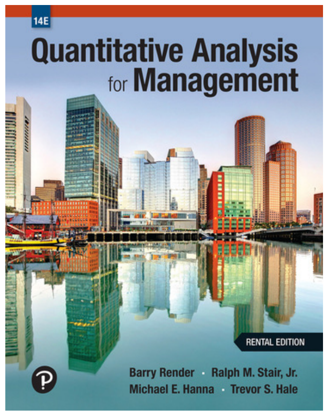
Chris J. Lloyd,
Data-Driven Business Decisions,
1st edition, Wiley, 2011
ISBN 0470619600



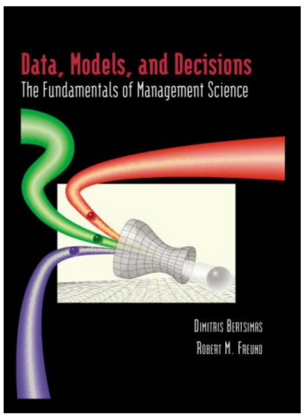
14 Investigating relationships between business variables
15 Describing the effect of a business input: Linear regression
16 The reliability of regression-based decisions
17 Multi-causal relationships and multiple regression
18 Product features, non-linear relationships and market segments.
19 Analysing data that is collected regularly over time.
20 Extending regression models – the sky is the limit

Bibliografía

Barry Render, Ralph M. Stair,
Michael E. Hanna & Trevor S. Hale
Quantitative Analysis for Management,
14th edition, Pearson, 2023
ISBN 0137943601
4 Regression Models



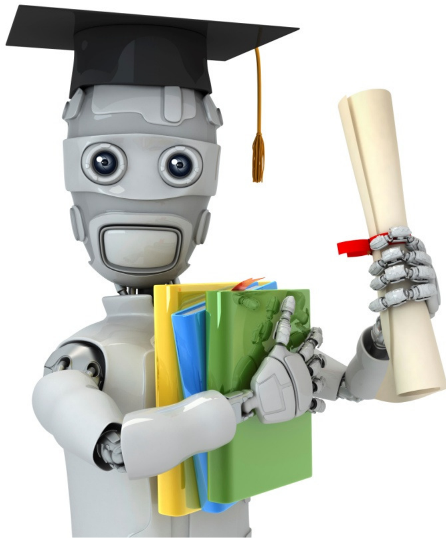
Dimitris Bertsimas & Robert Freund:
Data, Models, and Decisions:
The Fundamentals of
Management Science
Dynamic Ideas, 2004
ISBN 097591460X
6 Regression Models: Concepts and Practice



Bibliografía

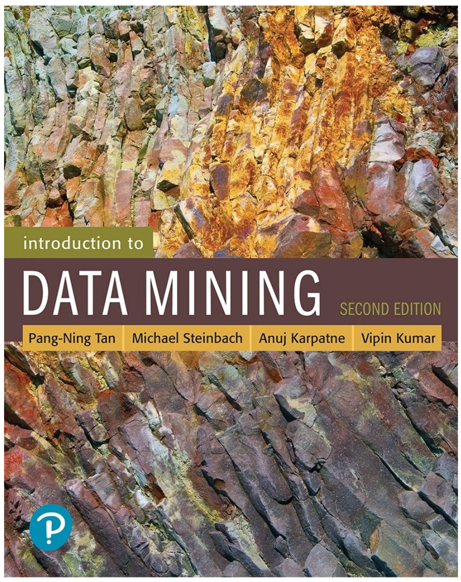
Andrew T. Ng:
Machine Learning
http://ml-class.org/

Originalmente (2011)
http://mlclass.stanford.edu/



Bibliografía

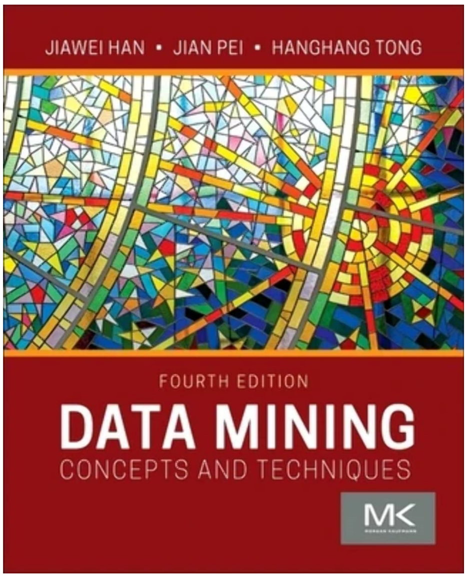
Pang-Ning Tan,
Michael Steinbach,
Vipin Kumar &
Anuj Karpatne:
Introduction to Data Mining,
2nd edition, Addison Wesley, 2018.
ISBN 0133128903



6.6 Logistic Regression

Bibliografía

Jiawei Han,
Jian Pei &
Hanghang Tong:
Data Mining:
Concepts and Techniques,
4th edition, Morgan Kaufmann, 2022.
ISBN 0128117605



6.5.1 Linear regression
6.5.3 Logistic regression