Maestría en Explotación de Datos y Descubrimiento del Conocimiento

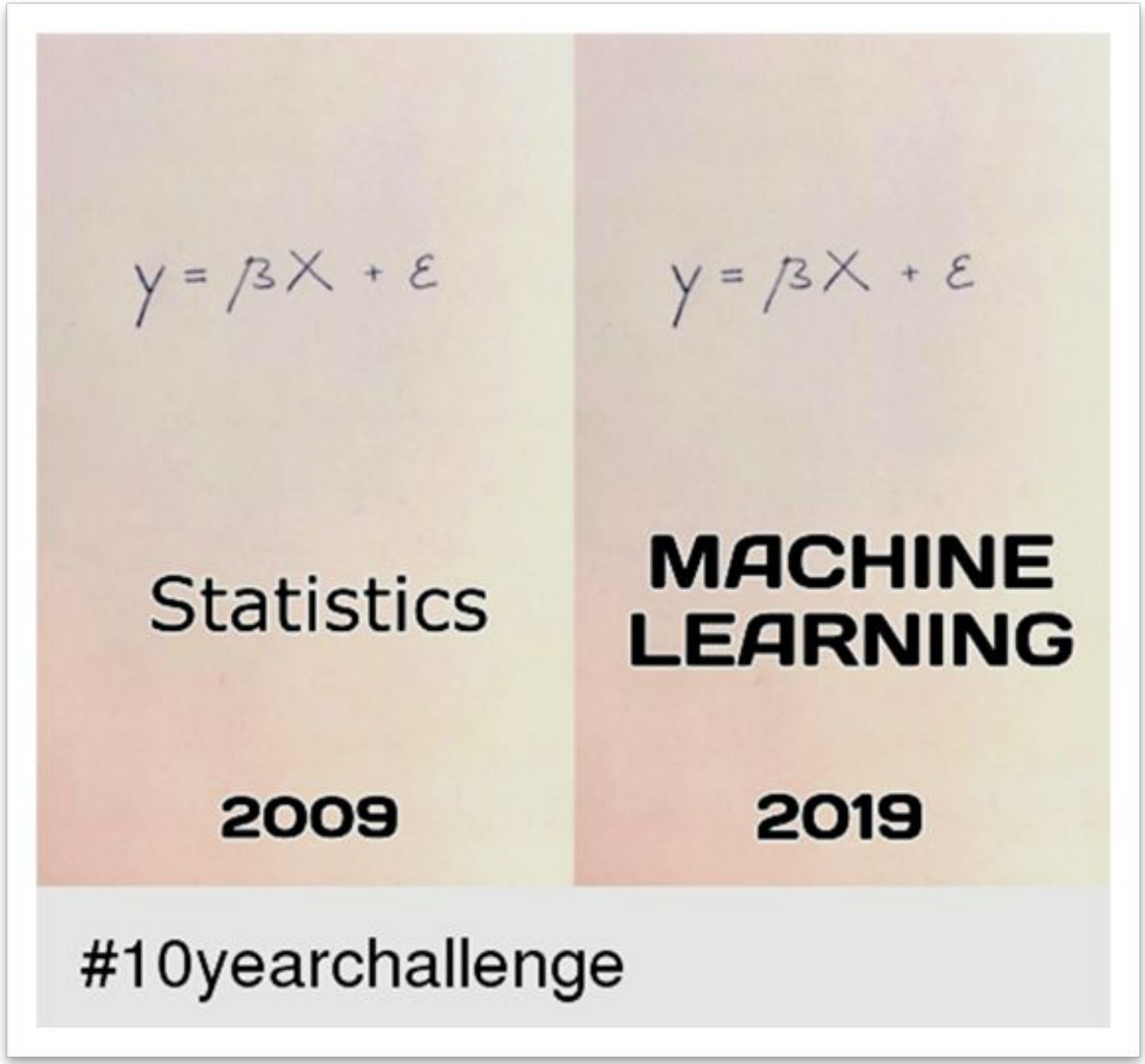
Unidad II, Clase 5

Aprendizaje Automático

Regresión

Laura de la Fuente, Hernán Bocaccio
Ayudantes: Gastón Bujía, Diego Onna y Sofía Morena del Pozo

Dirección de e-mail de la materia:
datawillconfess@gmail.com


$y = \beta X + \varepsilon$
Statistics
2009
$y = \beta X + \varepsilon$
MACHINE LEARNING
2019
#10yearchallenge

Itinerario de la clase

- Generalidades de regresión
- Regresión lineal simple
- Cuadrados mínimos (OLS)
- Estimadores de máxima verosimilitud (MLE)
- Regresión lineal múltiple
- Función costo
- Descenso por el gradiente
- Tipos de regularización
- Regresión con otros modelos clásicos

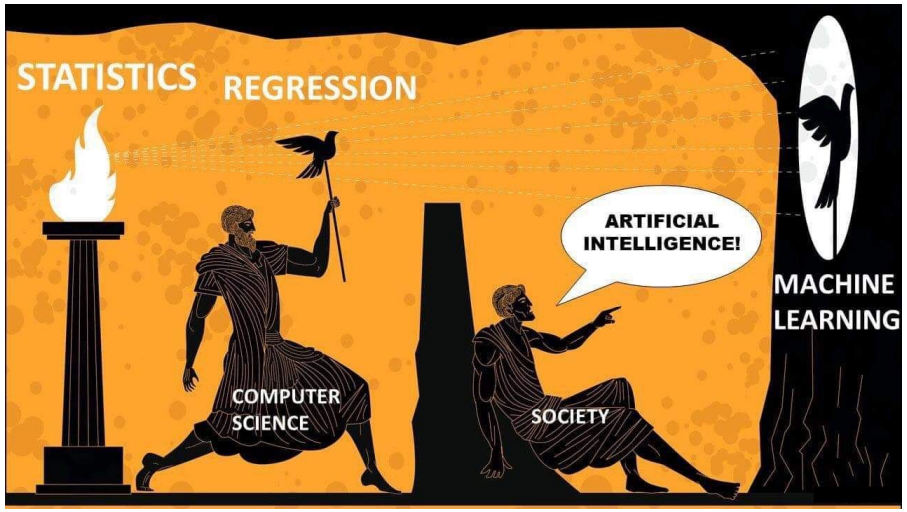
Regresión

Aprendizaje supervisado:

Los **datos están anotados** con la respuesta correcta que quiero predecir.

- Clasificación: Predecir una clase (variable categórica)
- **Regresión: Predecir un valor numérico (variable numérica)**

- **Herramientas sencillas** de análisis de datos para determinar y cuantificar **dependencias entre variables**
- Implementación de **modelos simples** que aplican a **gran variedad de problemas**
- Más allá de la utilidad de su implementación, su estudio permite comprender **conceptos fundamentales** que luego sientan las bases para otros análisis y **modelados más complejos**



Modelado

- Un modelo es una representación **simplificada** de un sistema y, en particular, de algún aspecto de interés del mismo (restringe las condiciones del sistema a estudiar). Todos incorrectos, algunos útiles.
- En general, el mismo se formula en el contexto de un **marco teórico** que lo sostiene (aunque también existen **modelos estrictamente empíricos**).
- El objetivo del modelo es **describir** el comportamiento del **sistema** y poder **predecirlo** en un futuro. El más simple es mejor (Parsimonia).

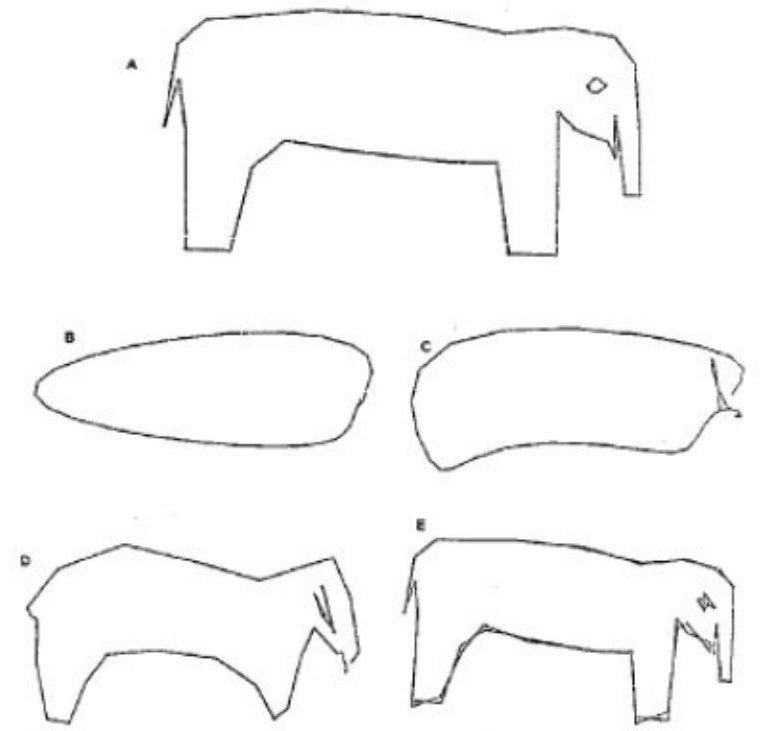


FIGURE 1.2. "How many parameters does does it take to fit an elephant?" was answered by Wel (1975). He started with an idealized drawing (A) defined by 36 points and used least squares Fourier sine series fits of the form $x(t) = \alpha_0 + \sum \alpha_i \sin(it\pi/36)$ and $y(t) = \beta_0 + \sum \beta_i \sin(it\pi/36)$ for $i = 1, \ldots, N$. He examined fits for $K = 5, 10, 20$, and 30 (shown in B–E) and stopped with the fit of a 30 term model. He concluded that the 30-term model "may not satisfy the third-grade art teacher, but would carry most chemical engineers into preliminary design."

Burnham, K. P., & Anderson, D. (2003). Model selection and multi-model inference. *A Pratical informatio-theoric approch. Sringer*.

Modelado

Una dada **magnitud** se puede modelar como una **función de muchas variables**. A mayor cantidad de variables, mayor es la **complejidad del modelado**.

+ Potencialmente una descripción más precisa y detallada
- Dificulta su planteo e interpretación
- Más propenso a sobreajuste y más sensibilidad al ruido (maldición de dimensionalidad)

Se busca un equilibrio entre la **simpleza del modelo** y la **efectividad** para **describir** el proceso/**sistema** en cuestión.

Modelo simple	Modelo complejo
$Y = f(x)$	$Y = f(x1, x2, x3, \ldots, xn)$

Si la respuesta es un valor se dice **univariado**, si son varios se dice **multivariado**

Tipos de dependencias

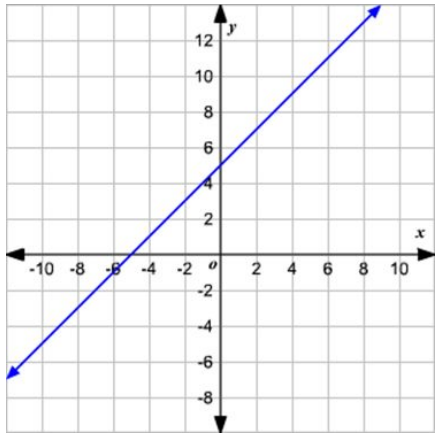
Modelo general lineal

$$Y_i = \beta_0 + \beta_1 X_i + ... + \varepsilon_i$$

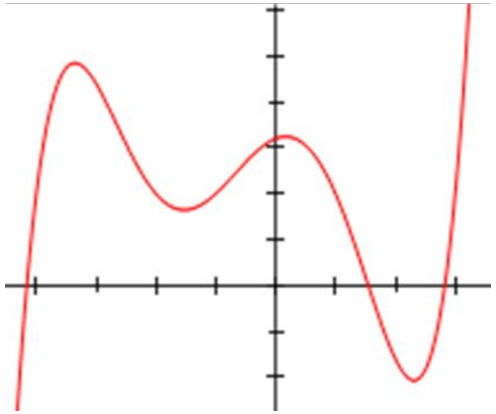
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^{\ 2} .... + \varepsilon_i$$

cuadrados mínimos / máxima verosimilitud

Función lineal



Función polinómica



La variable respuesta (VR) es una combinación lineal de las variables explicativas (VE) o de una transformación

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

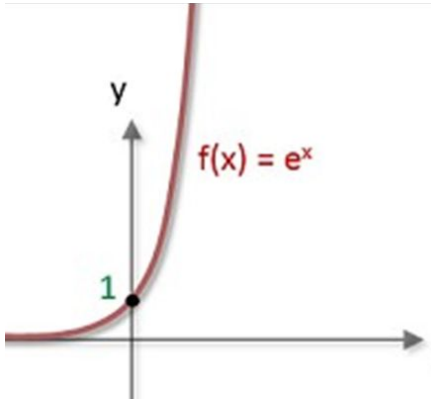
Modelo lineal generalizado

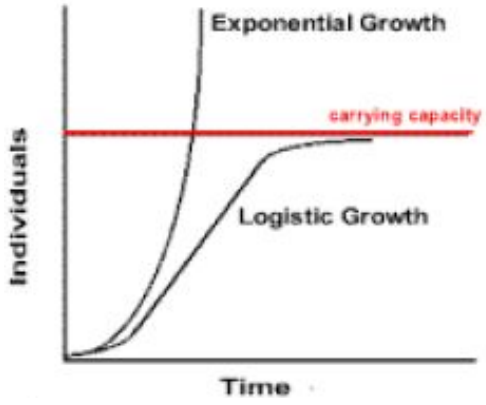
$$y = \beta_0 e^{\beta_1 X_i} + \varepsilon_i$$

máxima verosimilitud

Parámetros no lineales estimados iterativamente

Función exponencial

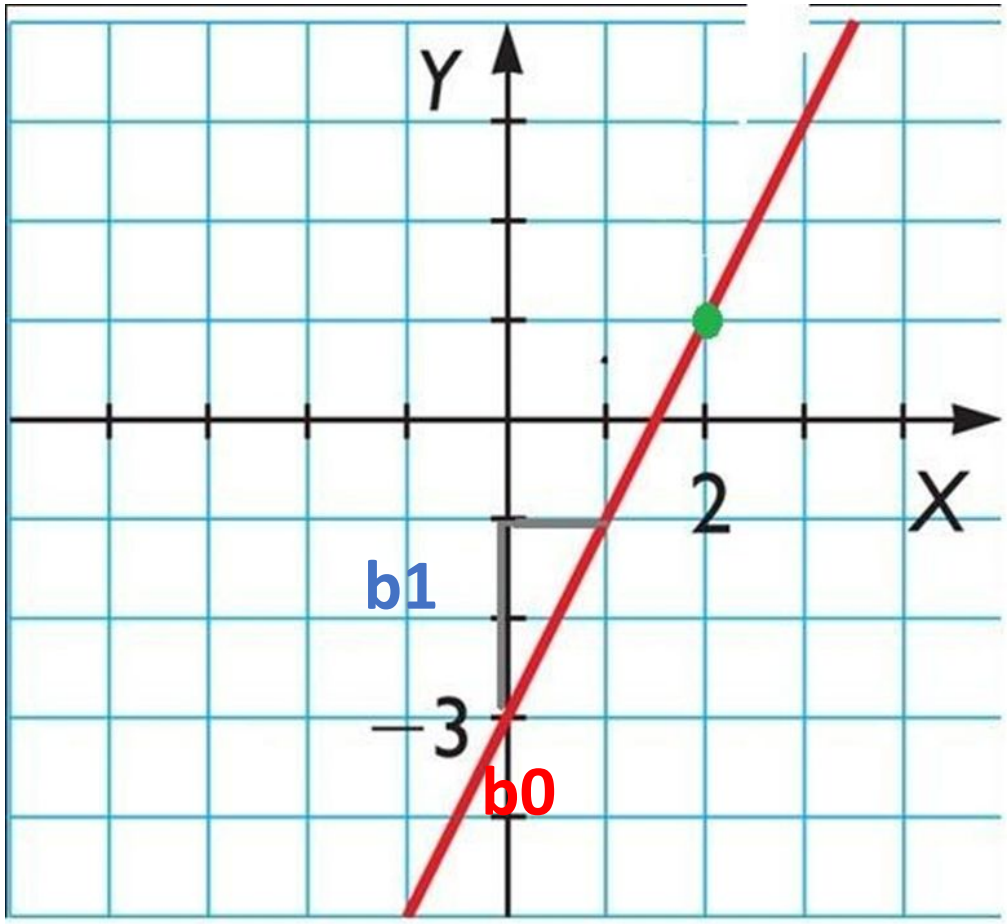




La VR es función no lineal de una combinación lineal de las VE o de una transformación

$$y(\mathbf{x}, \mathbf{w}) = f\left(w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) \right)$$

Regresión Lineal Simple



El objetivo de un **modelo de regresión lineal simple** es tratar de explicar la **relación** que existe entre una variable dependiente (variable respuesta) "Y" y una única variable explicativa X.

Respuesta — Parámetros — Atributos

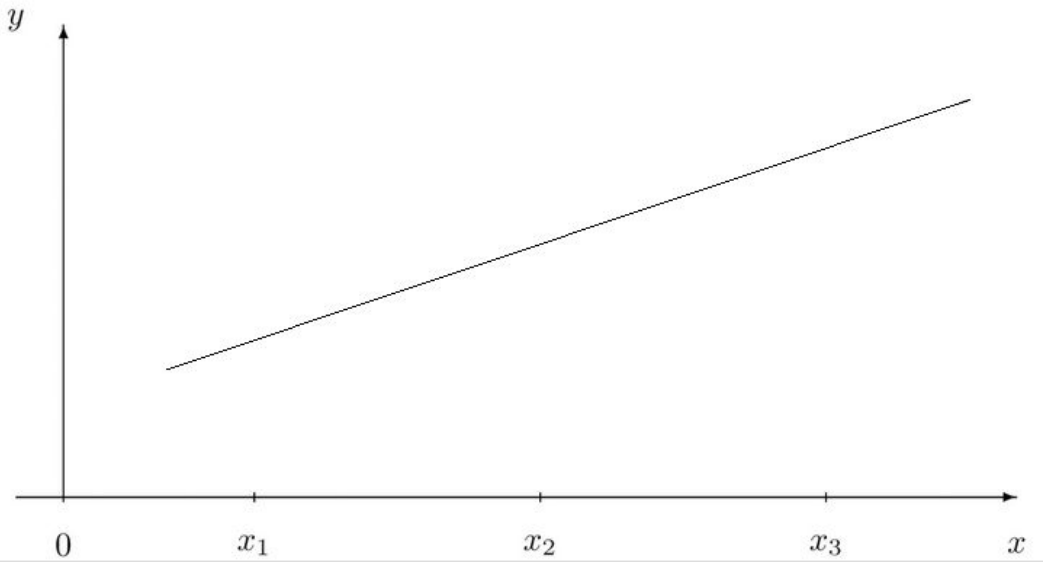
$$Y_i = \beta_0 + \beta_1 X_i$$

b0 = ordenada al origen o valor de la función cuando x=0

b1 = pendiente o tasa de cambio

Modelos determinísticos vs estadísticos

$$Y_i = \beta_0 + \beta_1 X_i$$

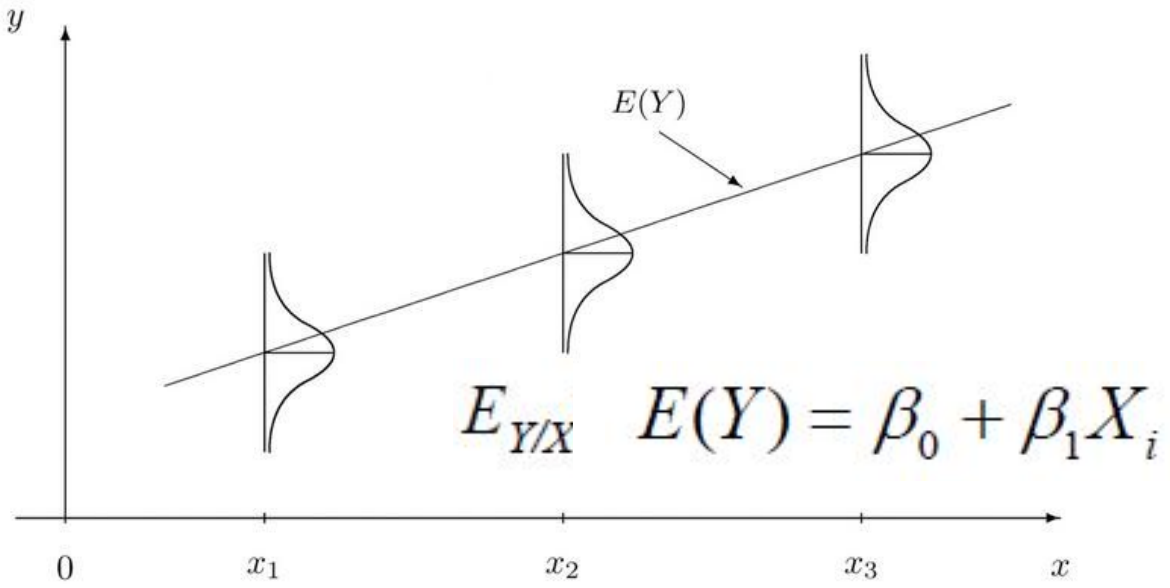


Modelo determinístico:

Dado un valor de X, Y queda determinada unívocamente

Componente aleatorio

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim NID\ (0, \sigma^2)$$



Modelo estadístico:

Dado un valor de X, la esperanza de Y queda determinada unívocamente. Existe variación aleatoria (error) que responde a distribución de probabilidades.

Regresión Lineal (enfoque estadístico)

Formulo un **modelo** que implique una **relación lineal** entre las variables en cuestión, pero además que tenga cierta **dispersión** respecto a dicha relación.

Esta dispersión puede representar diferentes efectos del sistema a medir como variables influyentes no tenidas en cuenta, ruido en los instrumentos de medición, etc.

equivalentes

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i=1\ldots n \qquad \varepsilon_i \sim NID\ (0,\ \sigma^2)$$

$$E_{Y/X} = \beta_0 + \beta_1 X_i$$

Valor esperado de $Y = \mu_{Y/x}$

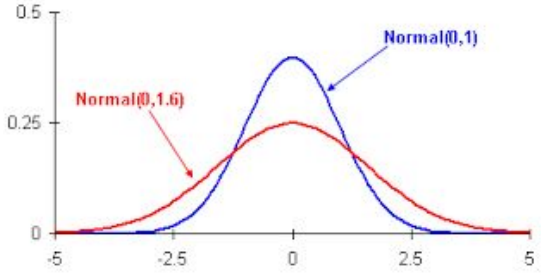
Beta 0 es la ordenada al origen (el valor que toma Y cuando X vale 0)

Beta 1 es la pendiente de la recta (e indica cómo cambia Y al incrementar X en una unidad)

Epsilon es una variable aleatoria cuya E(**ε**)=0 y Var(**ε**)=**σ²**. Esta variable incluye un conjunto grande de factores que afectan el valor esperado de Y. Se denomina error.

$$\varepsilon_i \sim NID\ (0,\ \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Regresión Lineal (enfoque estadístico)

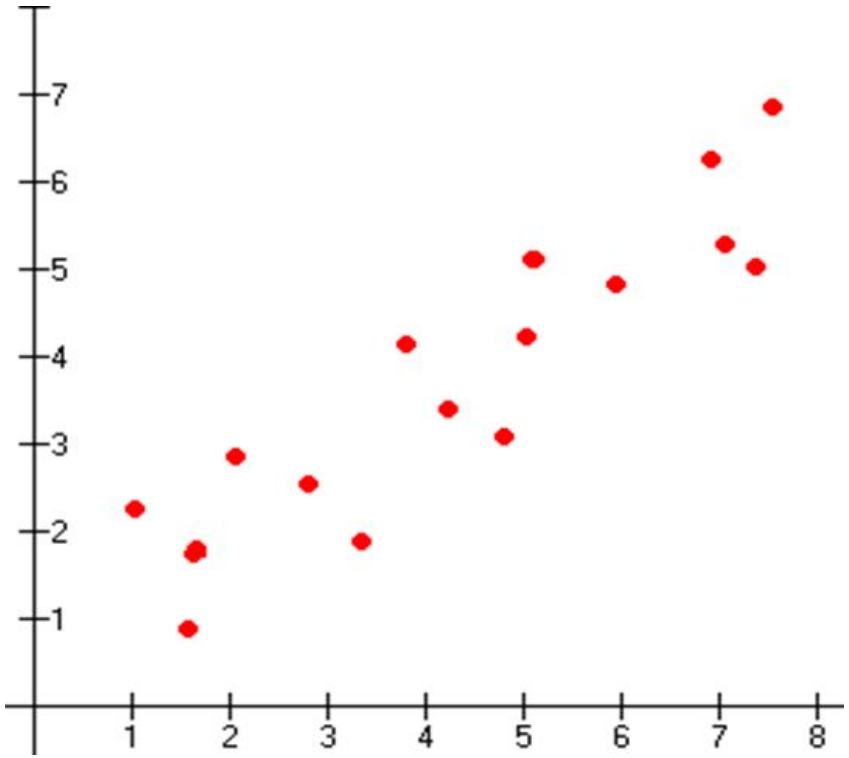
Dada la muestra que tengo, el objetivo es **determinar el modelo de regresión lineal** que **mejor describa a los datos**.

Diferentes parámetros definen **distintos modelos** y algunos de ellos representarán mejor que otros a la muestra.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Estamos interesados en modelar E(Y/X) = f (X);
pero observamos Y = f (X)

Para ello es conveniente **definir** algún **criterio** para **evaluar la precisión del modelo** respecto a la muestra que quiero modelar.

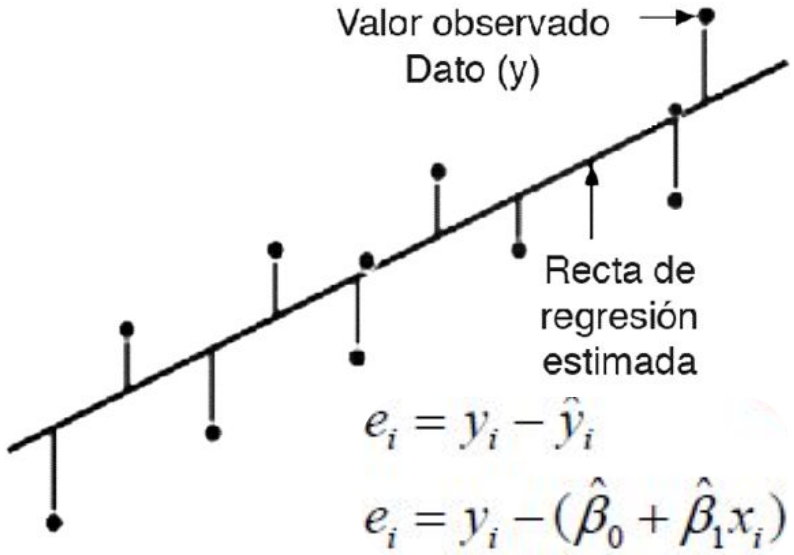


Error de predicción o residuo

¿Cuánto se “alejan” los datos de un determinado modelo?

Criterio de mínimos cuadrados

Para evaluar cuán preciso es un dado modelo, voy a cuantificar la suma de las distancias de cada elemento de la muestra respecto de la recta de regresión definida por dicho modelo.



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

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2$$

Error cuadrático medio

Mientras más pequeña sea la suma de distancias al cuadrado de cada dato respecto de la recta, mejor consideraré que es el modelo de regresión.

$\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores puntuales de β_0 y β_1

$$\sum_{i=1}^{n} e_i = 0$$

Mínimos cuadrados ordinarios (OLS)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2 \quad \longleftrightarrow \quad \mathrm{RSS} = \sum_{i=1}^{n} e^{(i)^2} = \sum_{i=1}^{n}\left(y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_1^{(i)}\right)^2$$

Es posible **minimizar esta suma** de distancias **analíticamente** de modo de hallar los **mejores estimadores** para los parámetros **β0 y β1** (método clásico).

Derivando e igualando a cero las derivadas parciales respecto de los parámetros, se buscan los mínimos de la función suma de **distancias al cuadrado**, y se puede demostrar que los valores para dichos parámetros serán:

$$\frac{\delta \mathrm{RSS}}{\delta \hat{\beta}_0} = 0 \qquad \frac{\delta \mathrm{RSS}}{\delta \hat{\beta}_1} = 0$$

Estimadores de parámetros

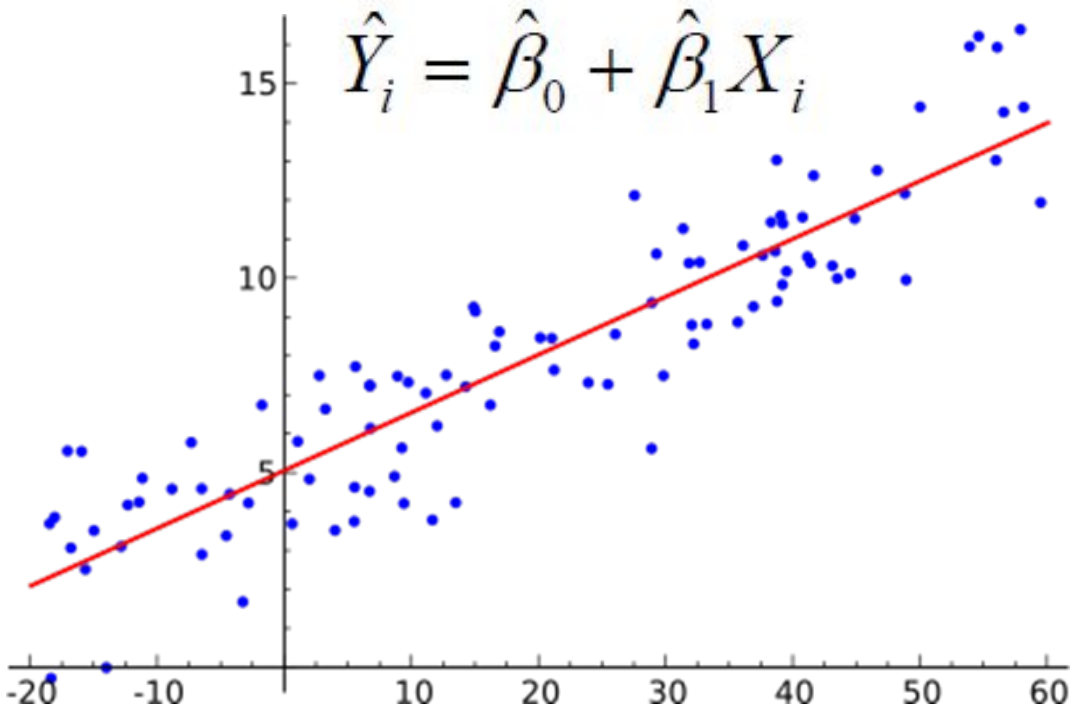
$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

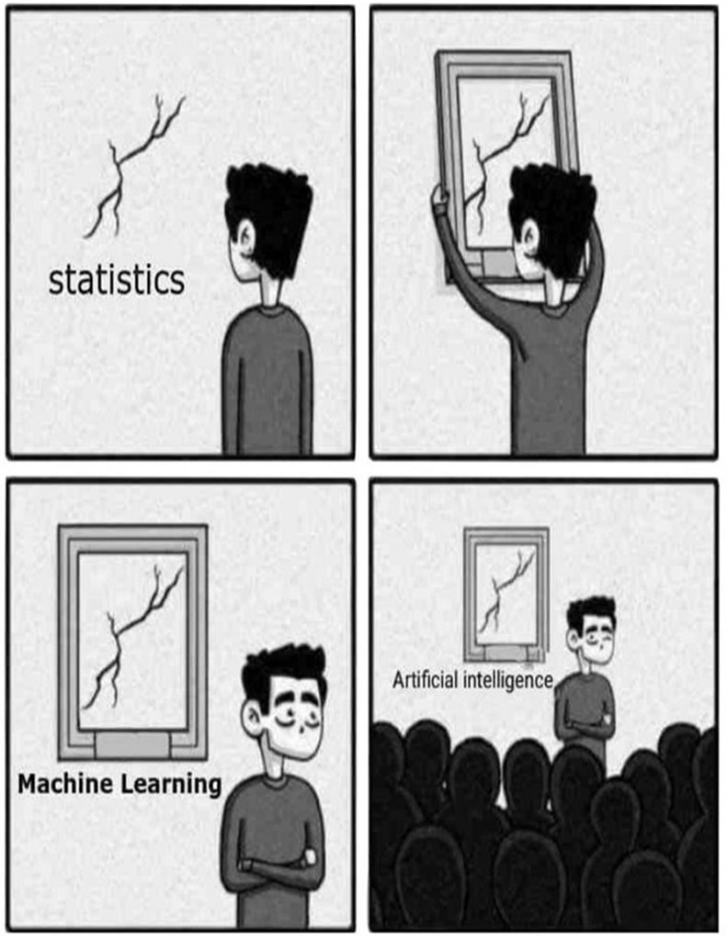
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Tienen una distribución de probabilidades

Mínimos cuadrados ordinarios (OLS)

Obtengo un modelo lineal que es el que mejor representa la dependencia entre las variables analizadas de mi muestra de datos.

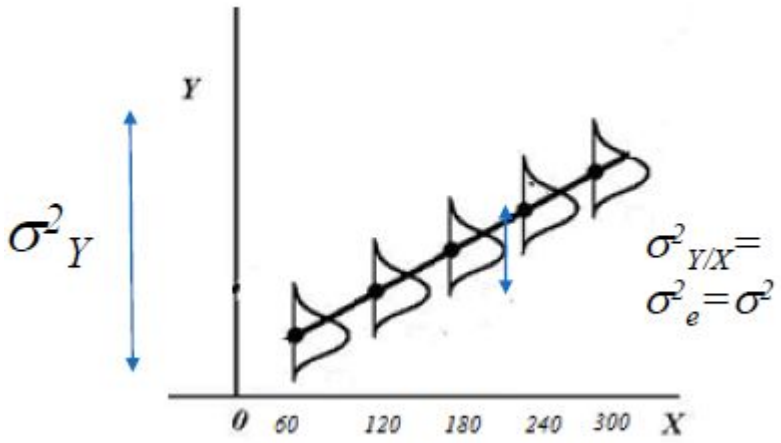




Dado este modelo, puedo predecir los valores que tendría Y, para un valor de X diferente de los que tomé en mi muestra.

Estimación de la varianza del modelo

Variabilidad total de los datos

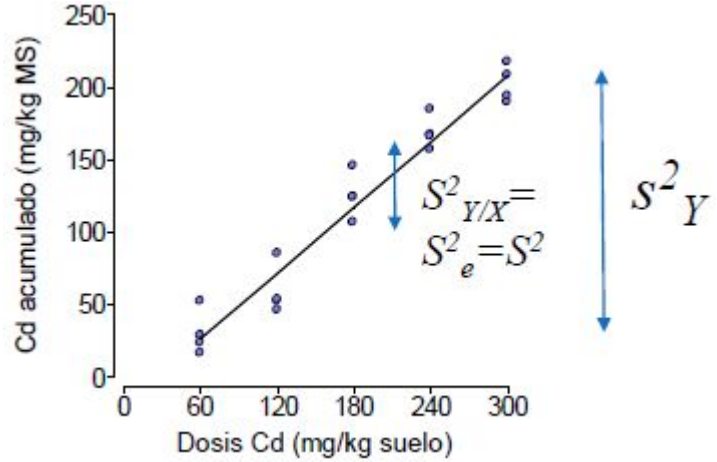


$$\sigma^2 = E\left[\left(Y_i - E_{Y/X}\right)^2\right]$$

$$\sigma^2 = E\left[\varepsilon_i^{\,2}\right]$$

$$\mathrm{TSS} = \sum_{i=1}^{n}\left(y^{(i)} - \overline{y}\right)^2$$

Variabilidad explicada por el modelo



$$s_e^2 = s^{2}_{\,Y/X} = CM_{error} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}{n-2} \cdots = \frac{\sum_{i=1}^{n} e_i^{\,2}}{n-2}$$

S_e se conoce como el EE del modelo o residual

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

variación total de VR = variación explicada por el modelo + variación no explicada

Coeficiente de determinación

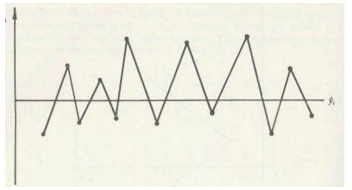
$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = \frac{SCexplicada}{SCtotal} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- El coeficiente de determinación puede interpretarse como la **proporción de variabilidad de Y que es explicada por X**.
- Es adimensional y toma valores entre 0 y 1
- Mide la proximidad de la recta ajustada a los valores observados de Y.
- Mientras más se acerque a 1, mejor representa el modelo a los datos.

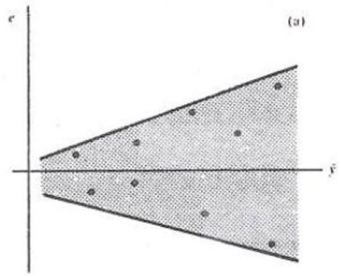
Supuestos

Para que el modelo pueda aplicarse y tengan sentido los resultados obtenidos y la significancia estadística, deben cumplirse ciertos supuestos sobre los datos (X sin error ie , > 10%):

1) Independencia: los residuos deben ser independientes entre sí.



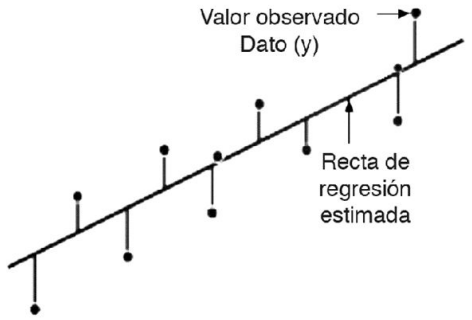
2) Homocedasticidad (igualdad de varianzas): Para cada valor de la variable X, la varianza de los residuos debe ser la misma (i.e. que el ajuste es igual de preciso independientemente de los valores de X).



Caso donde no se cumple

3) Esperanza de los errores igual a cero: Es decir, **E(ε)=0**.

4) Normalidad: para cada valor de la variable X, los residuos tienen distribución normal de media cero, es decir, **ε con distribución N(0,σ)**.



Máxima verosimilitud

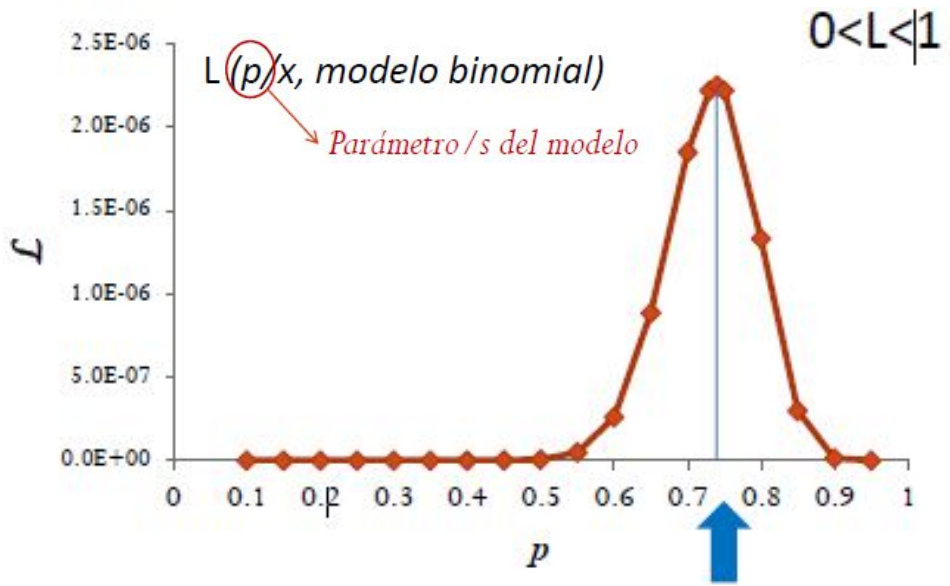
$$L(\theta / datos, modelo) \qquad \varepsilon_{ijk} \sim NID(0, \sigma^2)$$

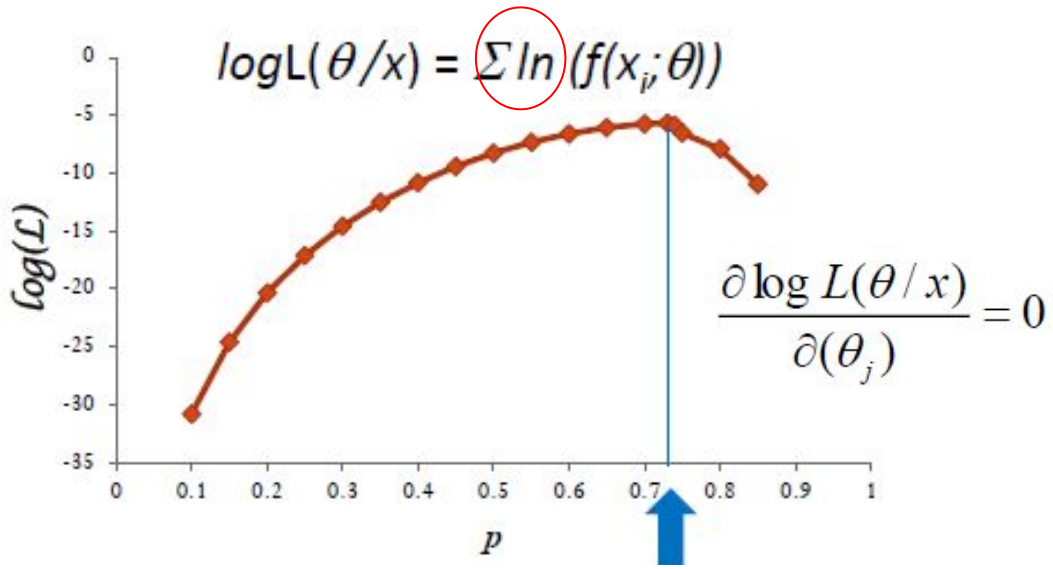
Estima los valores de los parámetros que hagan más verosímil (más probable) el resultado que hemos obtenido. Es decir, **que maximicen la probabilidad de obtener la muestra observada**.

Para ello usaremos la función de verosimilitud L, definida como la función de distribución conjunta, función de los parámetros desconocidos dada una muestra y una distribución de probabilidades.

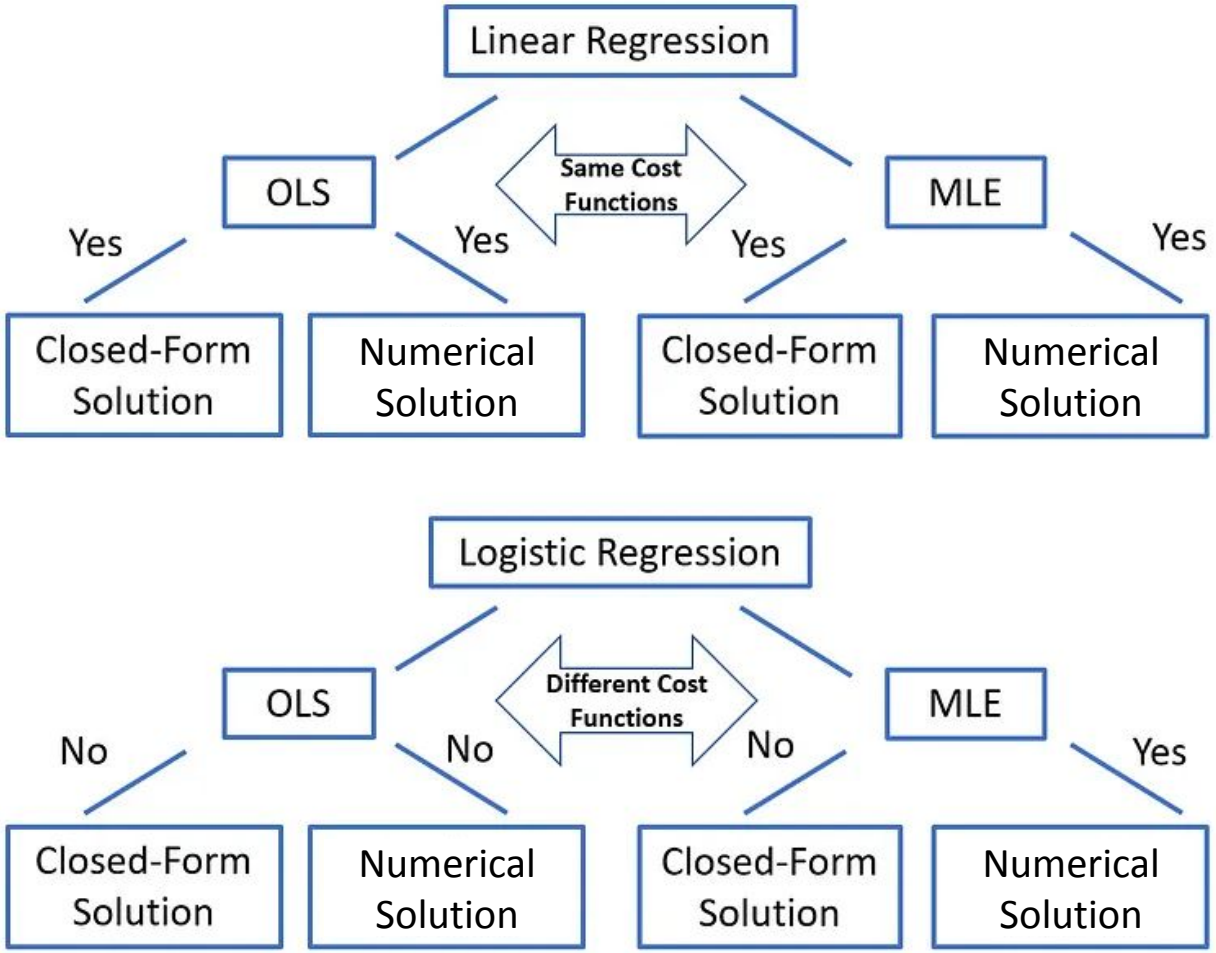
Elegiremos los valores de parámetros para los cuales la verosimilitud (o su log) sea máxima.

Si el tamaño de la muestra es grande, las estimaciones por MV proveen estimadores insesgados (la esperanza del estimador coincide con los parámetros) y consistentes (su varianza tiende a cero cuando n tiende a infinito).





Máxima verosimilitud



- Regresión lineal, admite la solución analítica por OLS, y es idéntica a MLE.
- Función costo de cuadrados mínimos es la que maximiza la verosimilitud.
- También métodos numéricos iterativos.
- Para otras regresiones, ya no sirve OLS (función costo no convexa).
- Se usa MLE, no siempre tiene solución analítica, pero sí numérica.
- Defino función costo a minimizar que maximize verosimilitud como -log(L) y uso método numérico iterativo.
  [Ej. regresión logística y log loss]

$$-\log(\text{likelihood}) = \text{loss function}$$

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

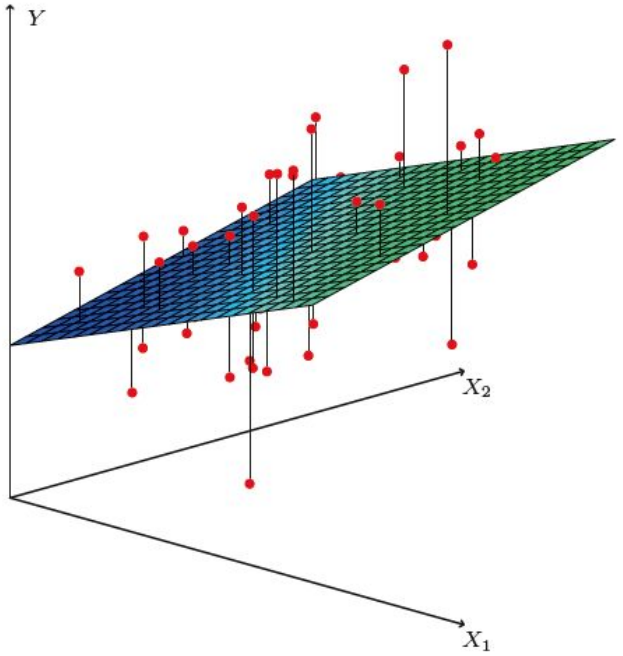
$$J(\theta) = -\frac{1}{m} \sum \left[y^{(i)} \log(h\theta(x(i))) + \left(1 - y^{(i)}\right) \log(1 - h\theta(x(i))) \right]$$

Regresión Lineal Múltiple

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Álgebra →

$$\hat{Y} = X^T \hat{\beta}$$
$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

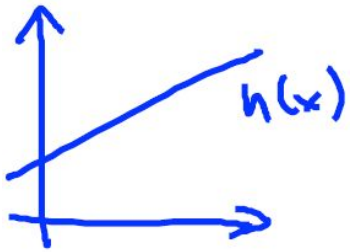


- Hiperplano de ajuste en el hiperespacio p-dimensional.
- Puedo resolver analíticamente con OLS, y análogamente con MLE, pero tengo muchos parámetros (complicado).
- Puedo resolver algebraicamente, pero tengo problemas al calcular la matriz inversa (correlaciones de atributos).
- Puedo resolver numéricamente MLE, pero ya que son muchos parámetros, tengo que definir una forma óptima de minimizar la función costo iterativamente. Para esto, un algoritmo muy usado es …**el descenso por el gradiente.**

De manera general, cualquier problema de regresión se resuelve con MLE, minimizando la función costo asociada iterativamente a partir del descenso por el gradiente.

Descenso por el gradiente

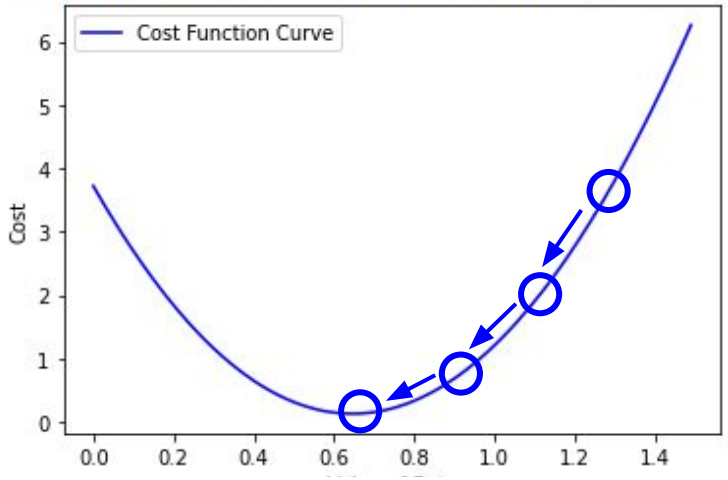
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$



Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$



<u>Proceso iterativo:</u>

- Empiezo por valores arbitrarios de parámetros
- Actualizo cada parámetro moviendome un poco en la dirección en la que el gradiente desciende
- Repito el proceso hasta convergencia

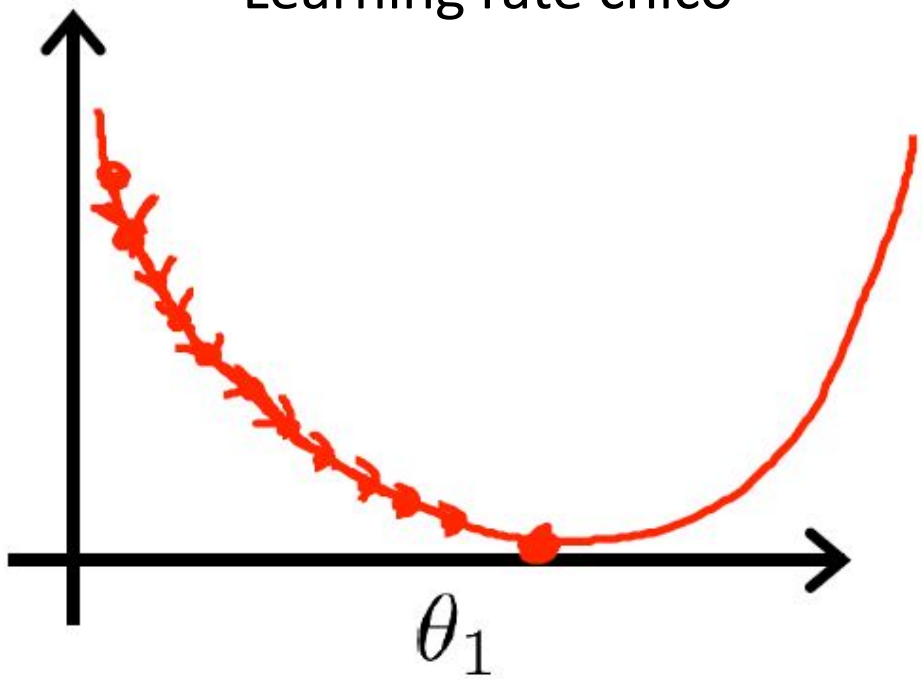
$$\text{repeat until convergence } \{$$
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$
$$(\text{for } j = 1 \text{ and } j = 0)$$
$$\}$$

Descenso por el gradiente

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

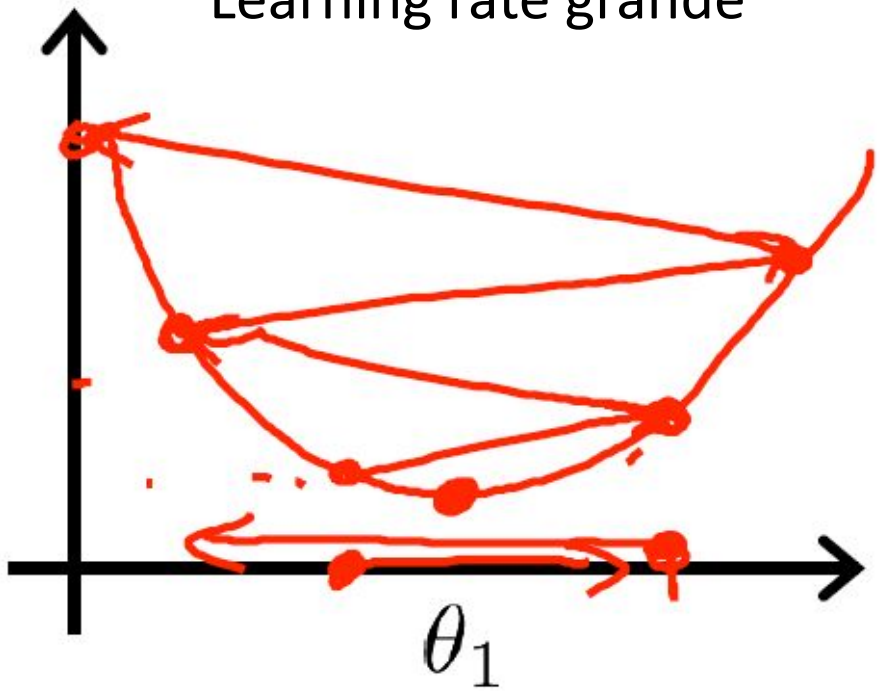
Learning rate (alpha; lambda): Factor que aporta al paso en el que me muevo cada iteración en la dirección que desciende el gradiente

Learning rate chico



Tardo en encontrar el mínimo, a riesgo de cortar la iteración antes de encontrar un valor suficientemente óptimo

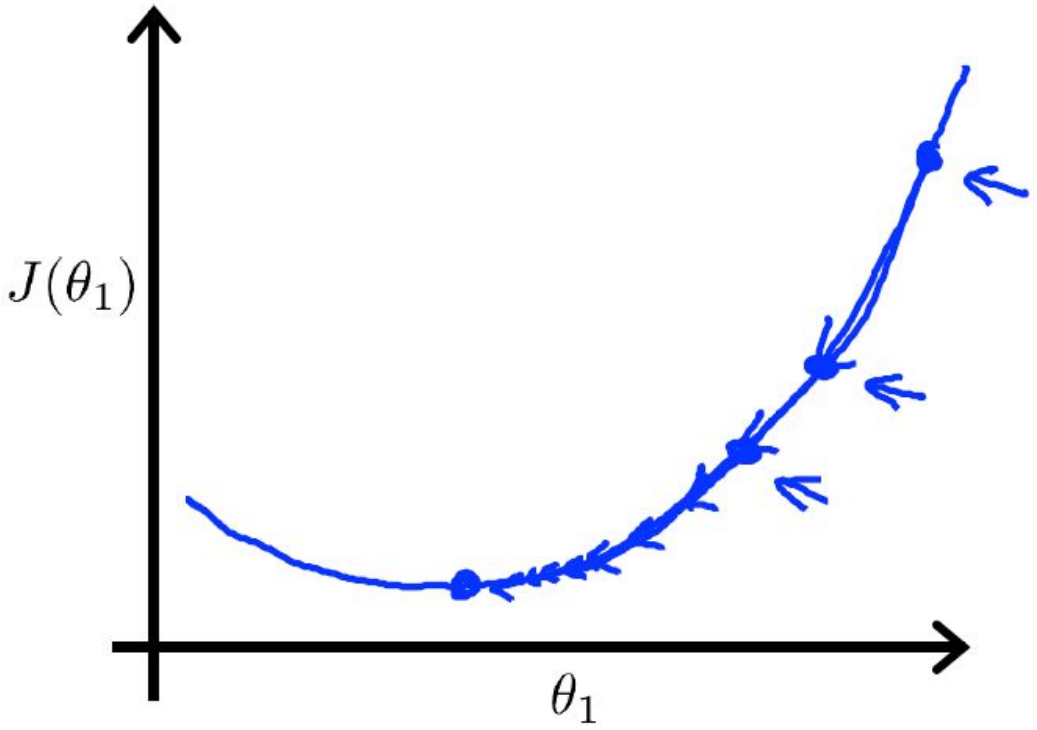
Learning rate grande



Pasos muy grandes me pueden hacer oscilar alrededor del mínimo a riesgo de nunca encontrarlo

Descenso por el gradiente

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$



El tamaño del paso también está determinado por la amplitud del gradiente evaluado en los valores de parámetros antes de moverme.

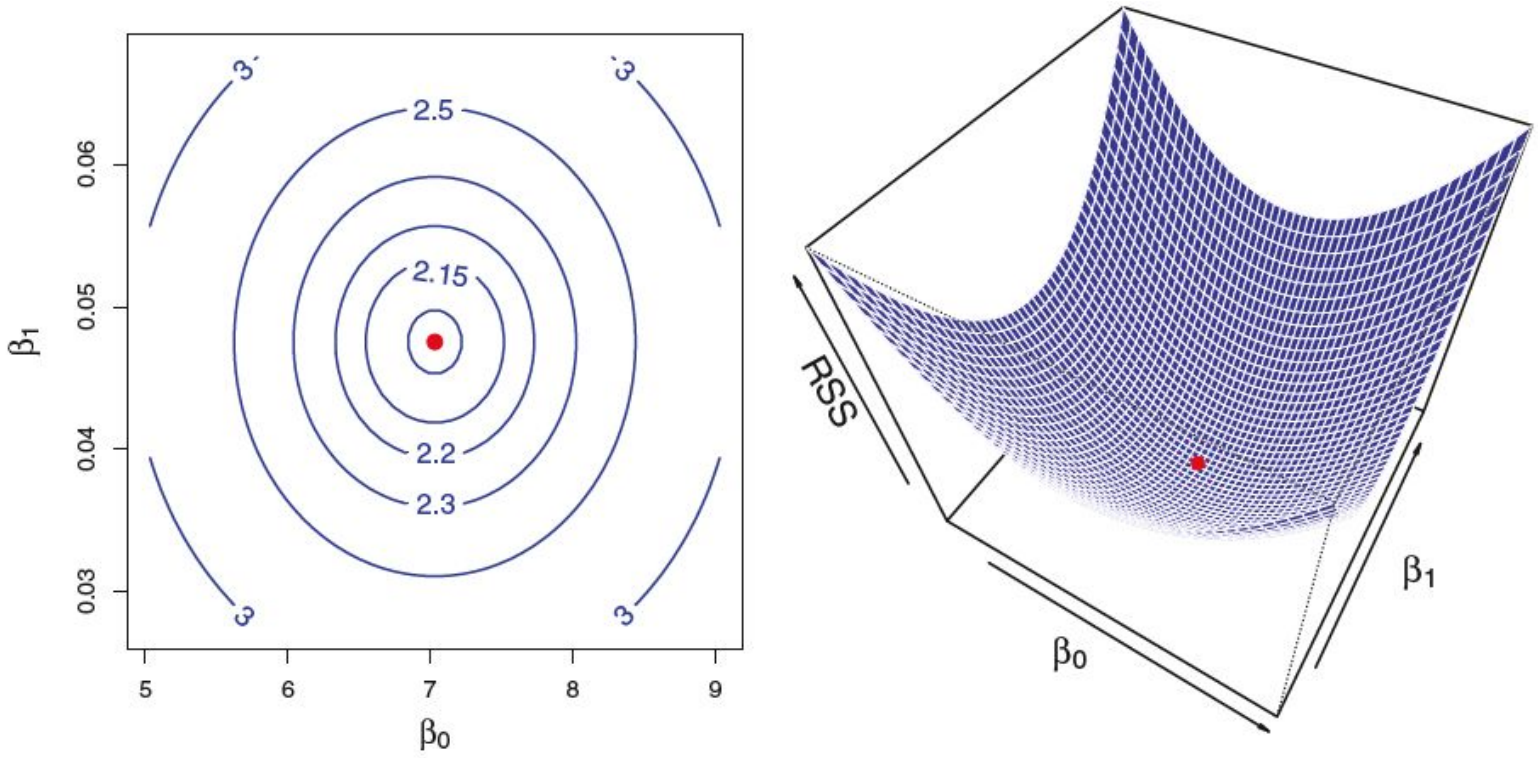
Cuando me acerco al mínimo, la derivada se hace más chica, y el paso también, aún cuando el learning rate esté fijo.

No es necesario ir reduciendo el alpha.

Función costo

$$\text{RSS} = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 \quad \longleftrightarrow \quad \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

Igual a minimizar

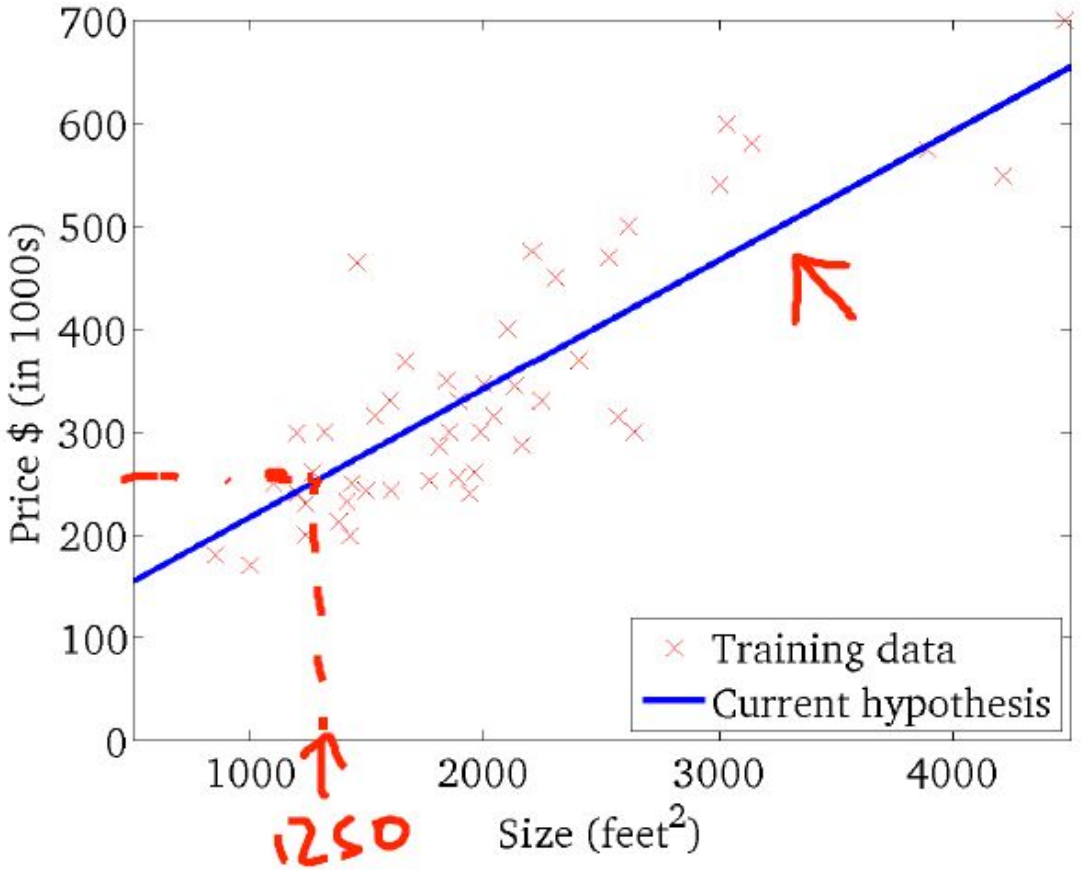


En el espacio de parámetros, combinaciones de parámetros dan el **mismo valor de error** y determinan **contornos de error**. Para RSS son **elipses**, según cuánto cambian el error cada parámetro.

Descenso por el gradiente

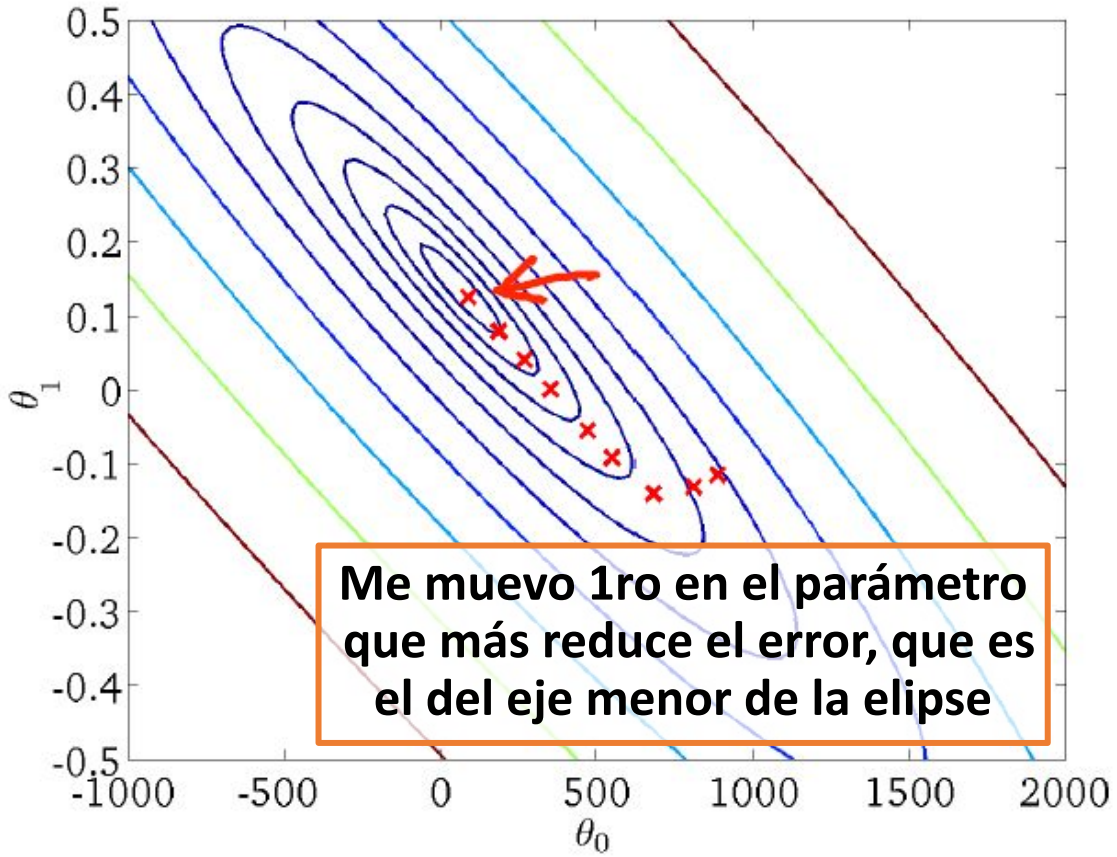
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



Descenso por el gradiente

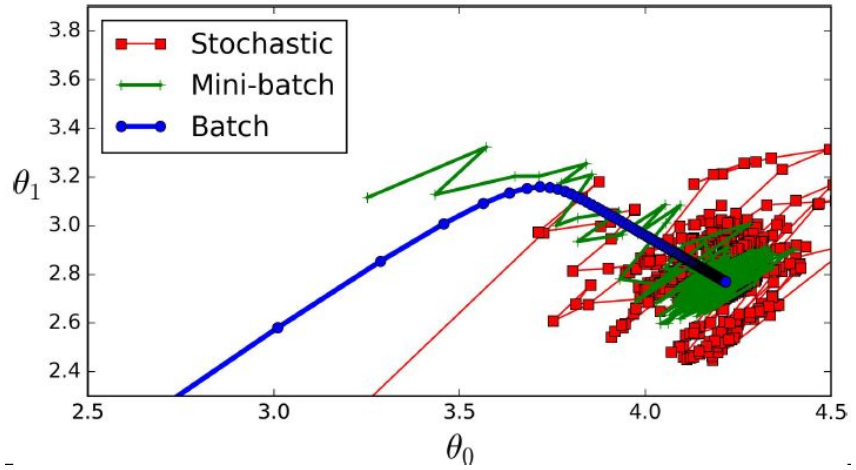
Actualizo los pesos en cada iteración, pero no necesariamente con todos los datos

- “Batch” Gradient Descent:

  Usa todos los datos de entrenamiento para calcular la función costo y el gradiente.
- “Stocastic” Gradient Descent:

  Usa una única instancia de los datos de entrenamiento, elegida al azar, para calcular la función costo y el gradiente.
- “Mini-batch” Gradient Descent:

  Usa un subconjunto de datos de entrenamiento para calcular la función costo y el gradiente.



Descenso por el gradiente

- Necesito definir el learning rate

+ No necesito despejar parámetros en ecuaciones muy complicadas como pasa cuando MLE admite solución analítica

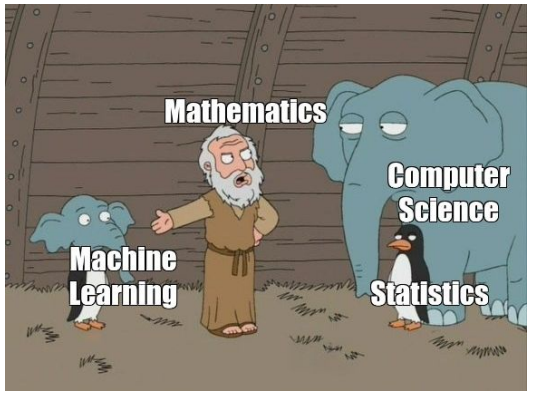
+ Me permite encontrar MLE, aún cuando no admiten solución analítica

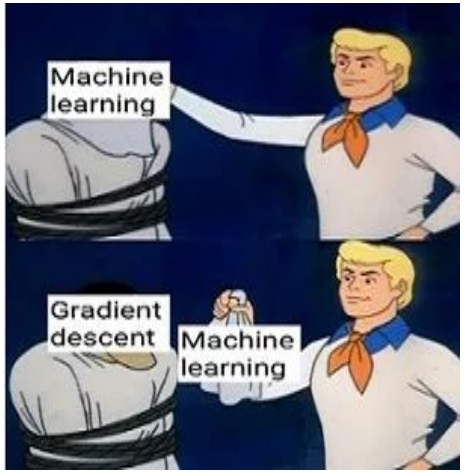
- Necesito hacer muchas iteraciones

+ Son menos iteraciones que si no tuviese un criterio óptimo de cómo actualizar los parámetros

+ No necesito calcular inversas de matrices que a veces no están bien condicionadas

+ Funciona bien, aún cuando el número de instancias es muy grande





Regularización

$$\text{RSS} = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 + \lambda \sum_{j=1}^{M} \left|\hat{\beta}_j\right|^q$$

Agregamos un término de penalización en la función costo

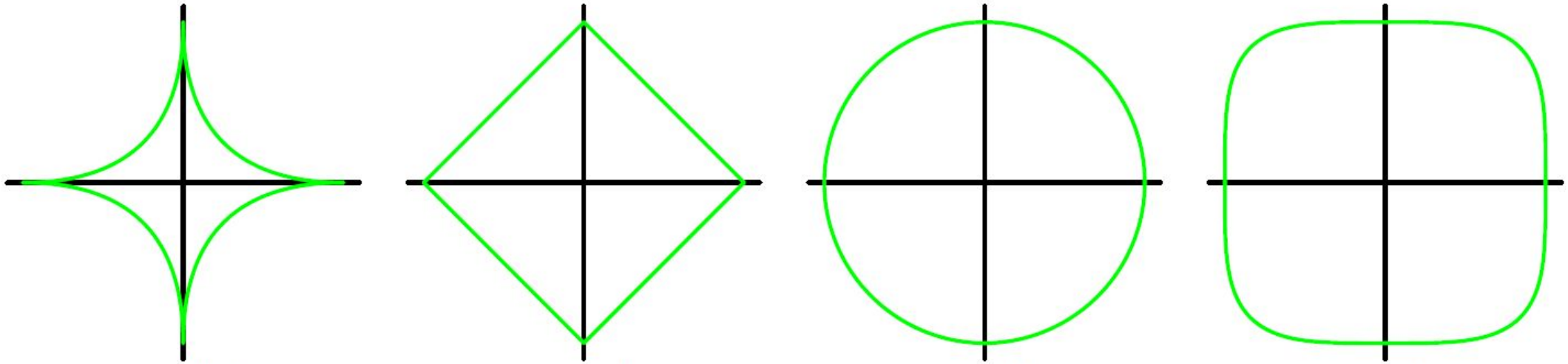
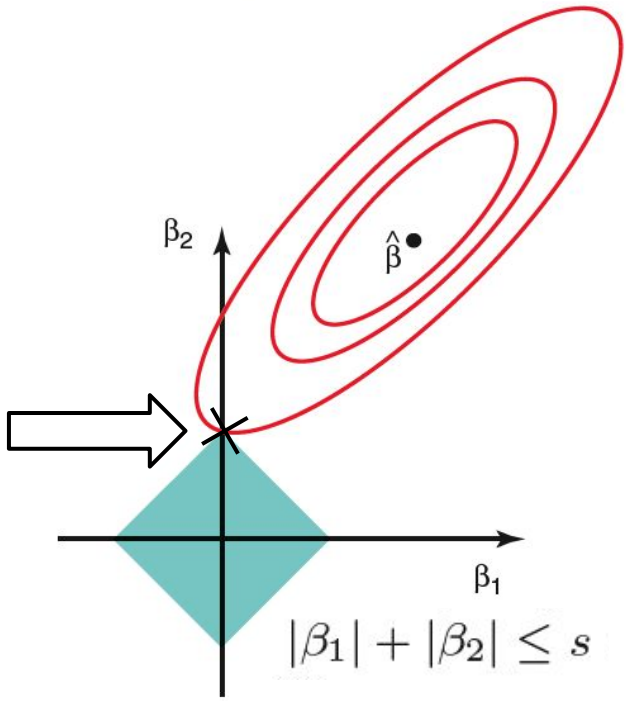
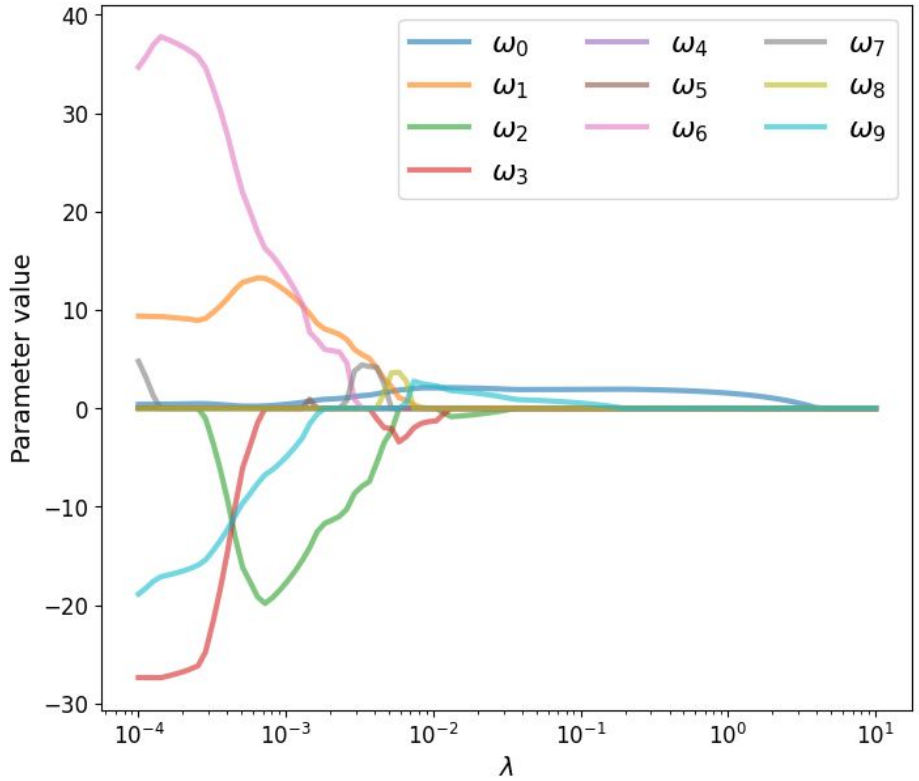


Figure 3.3 Contours of the regularization term in (3.29) for various values of the parameter q.

Regularización: Lasso

$$E_{\text{lasso}}(\boldsymbol{\omega};\lambda) = \frac{1}{2}\sum_{i=1}^{N}\{y(x_i,\boldsymbol{\omega}) - t_i\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M}|\omega_i|$$



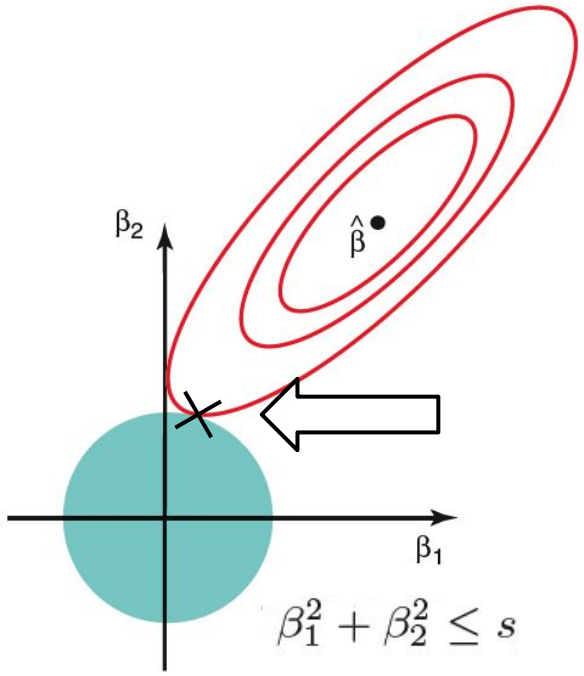


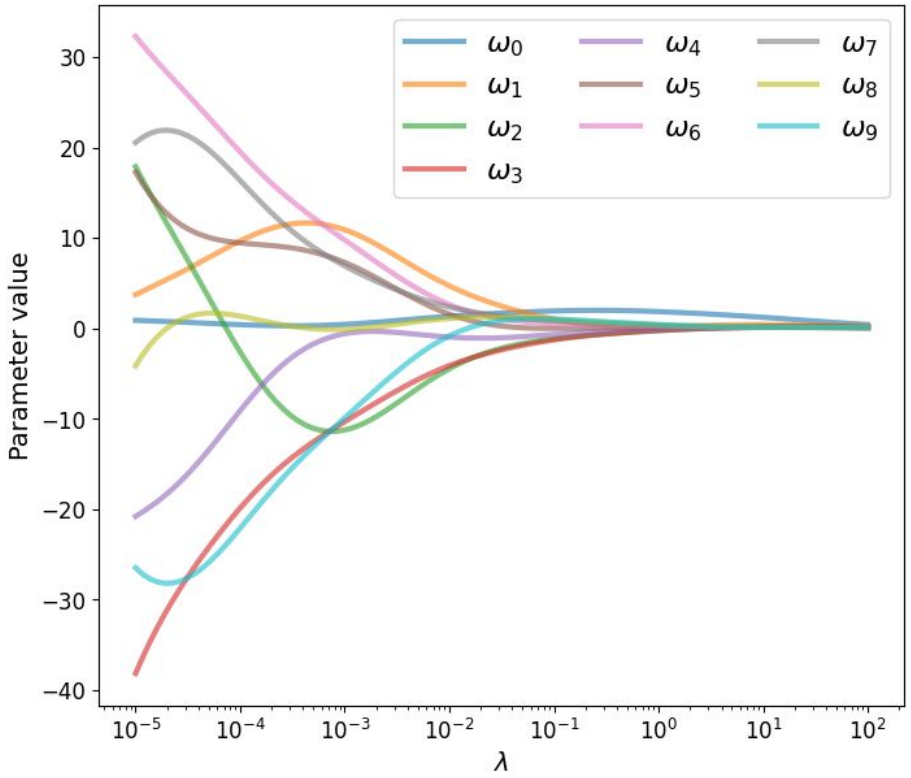
<u>Lasso o l1 (q=1):</u>

Cuando penalizo, es posible que parámetros se hagan cero (feature selection)

Regularización: Ridge

$$E_{\text{ridge}}(\boldsymbol{\omega};\lambda) = \frac{1}{2}\sum_{i=1}^{N}\{y(x_i,\boldsymbol{\omega}) - t_i\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M}\omega_i^2$$





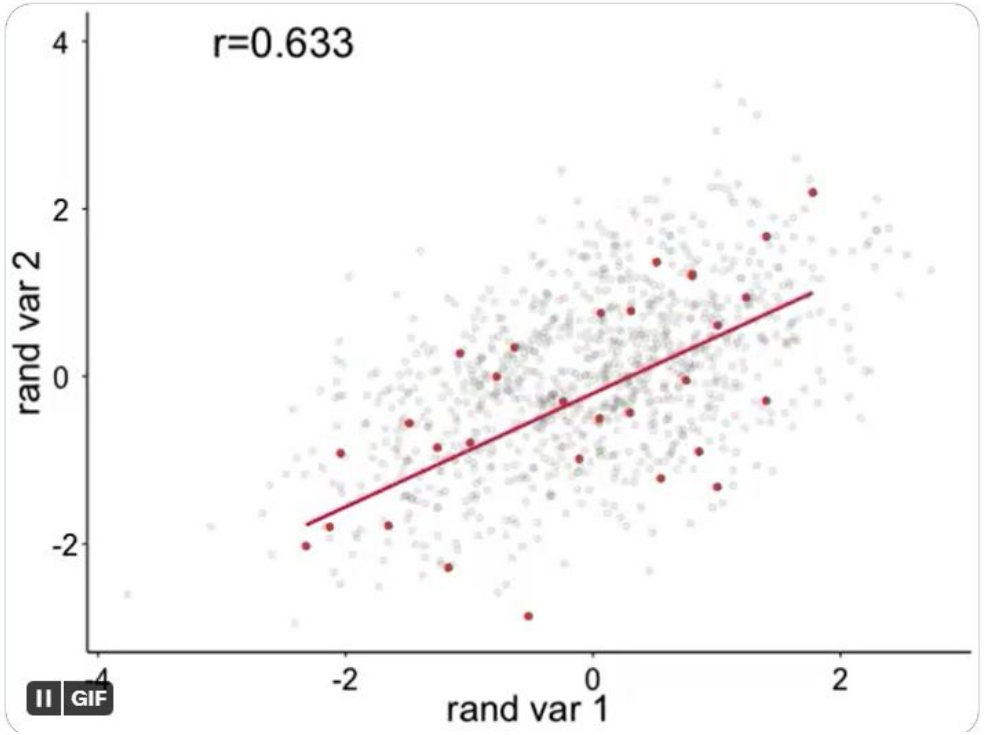
<u>Ridge o l2 (q=2):</u>

Cuando penalizo, parámetros se pueden hacer muy chicos pero distintos de cero

<u>Elastic-Net:</u> Combinación de Lasso y Ridge

Variaciones de sampleo y validación cruzada

Si uso subconjuntos de datos, el ajuste cambia. La validación cruzada le da robustez a la estimación del modelo, en contraste con ajustar todos los datos sin considerar que estoy submuestreando.



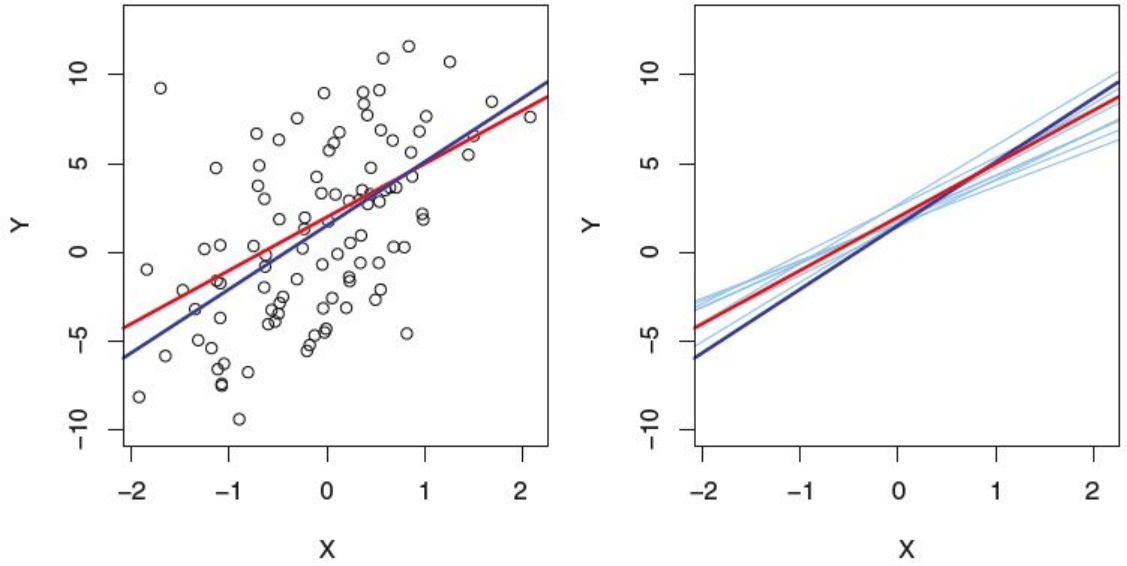


FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship,* $f(X) = 2 + 3X$, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for* $f(X)$ *based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Regresión con otros modelos (clásicos)

KNN regresión: Dada una nueva instancia, devolver el promedio (ponderado) de los valores de sus vecinos

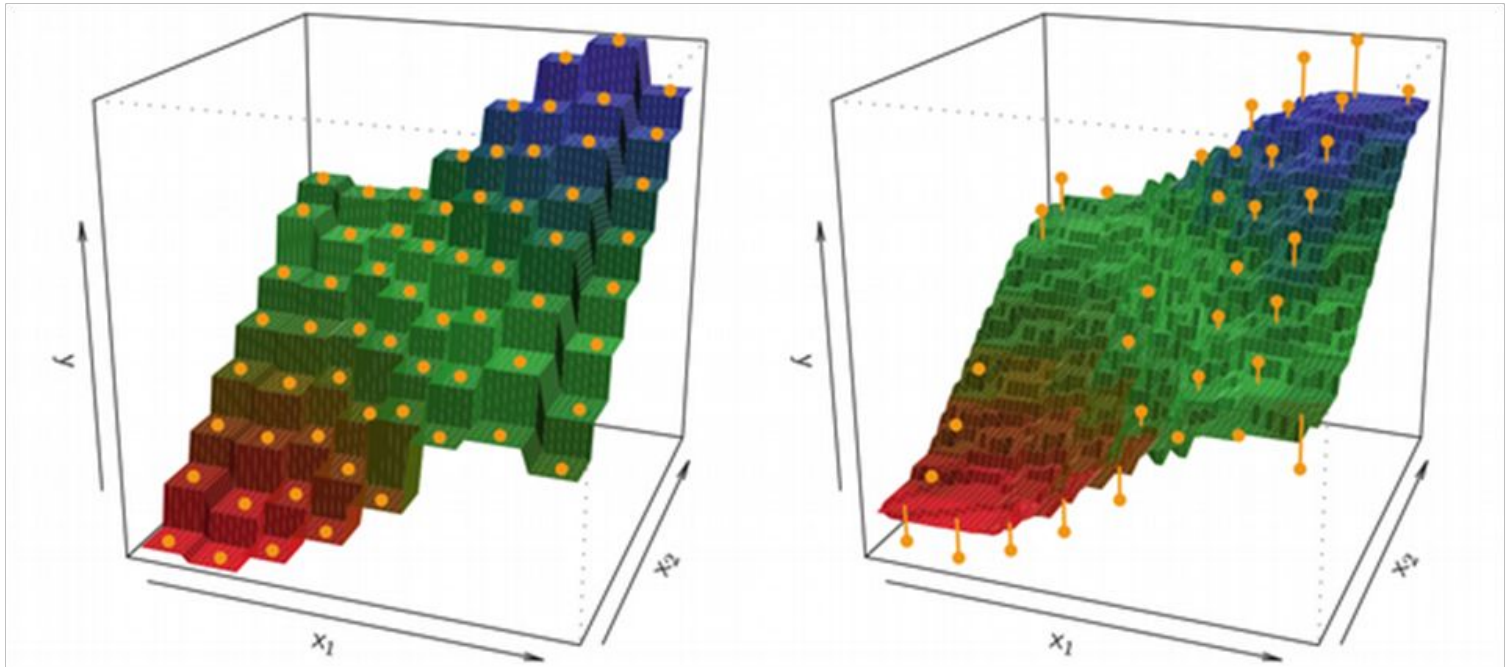
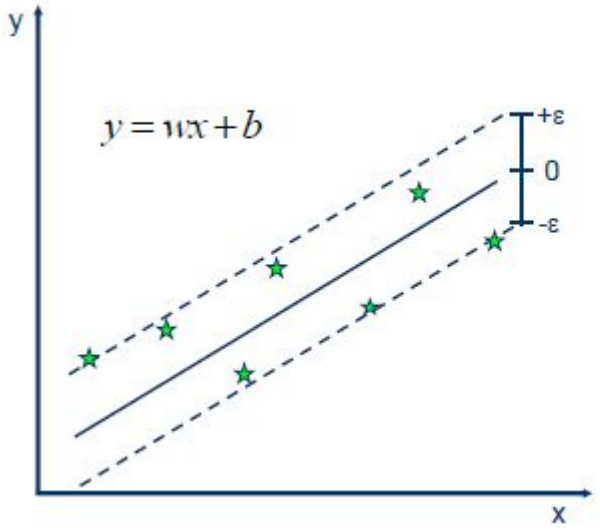


FIGURE 3.16. *Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots).* Left: $K = 1$ *results in a rough step function fit.* Right: $K = 9$ *produces a much smoother fit.*

Regresión con otros modelos (clásicos)

SVM regresión (SVR): También busco un hiperplano, pero ahora defino un márgen de tolerancia (epsilon), de manera tal que los datos estén dentro de ese márgen. Puedo también flexibilizar para reducir efectos de outliers.

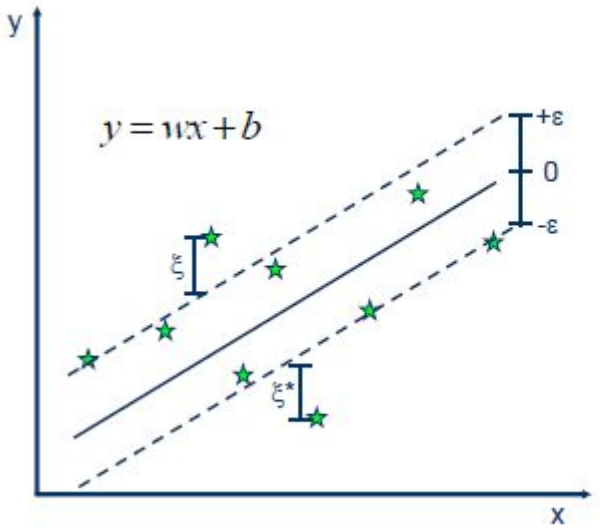


- Solution:

$$\min \frac{1}{2}\|w\|^2$$

- Constraints:

$$y_i - wx_i - b \le \varepsilon$$
$$wx_i + b - y_i \le \varepsilon$$



- Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)$$

- Constraints:

$$y_i - wx_i - b \le \varepsilon + \xi_i$$
$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0$$

Huyamos hacia los Colabs…!!