La recta de regresión

Presentación del problema. Diagrama de dispersión. El error cuadrático medio. Coeficientes de la recta.

Grado de Biología sanitaria

M. Marvá

e-mail: marcos.marva@uah.es

Unidad docente de Matemáticas, Universidad de Alcalá

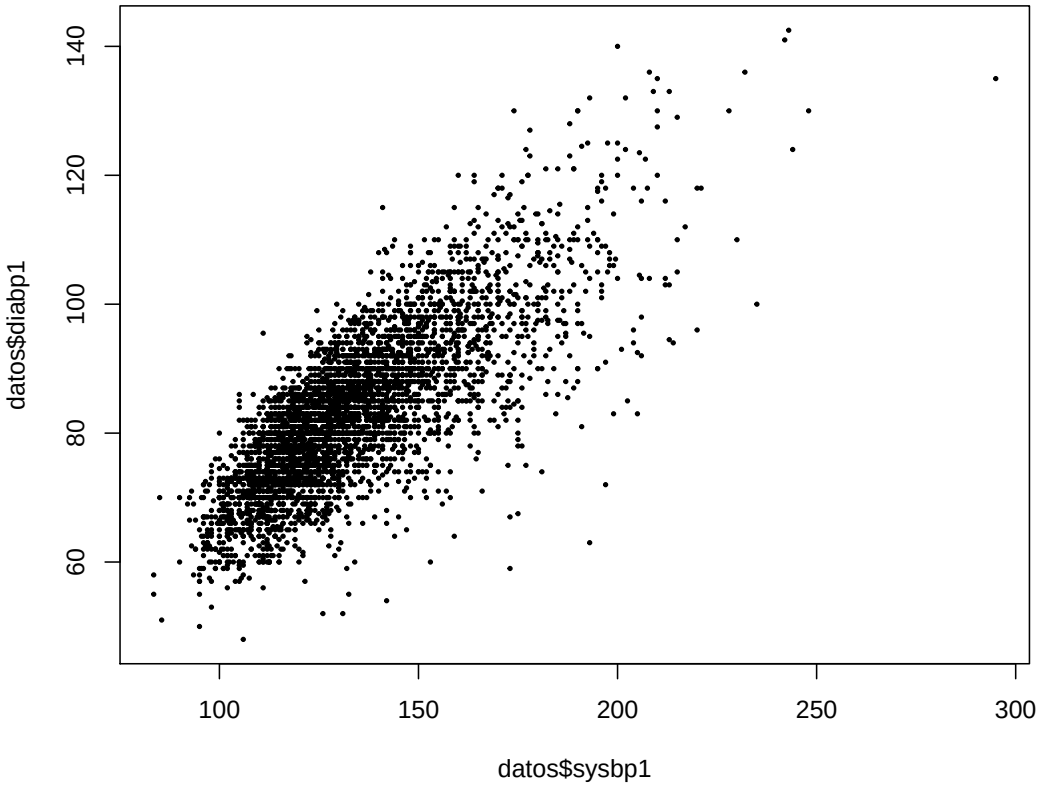
13 de septiembre de 2017

El problema

Trabajamos con los datos del estudio de `Framinham`

- ¿Hay relación entre las tensiones sistólica y diastólica de los individuos?

REPRESENTAR LOS DATOS

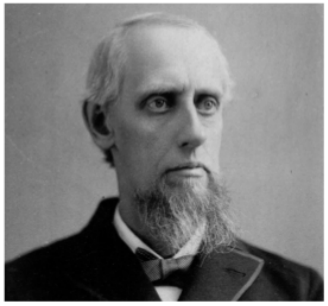


El problema

- Número de “cricks” que hace un grillo y la temperatura ambiente

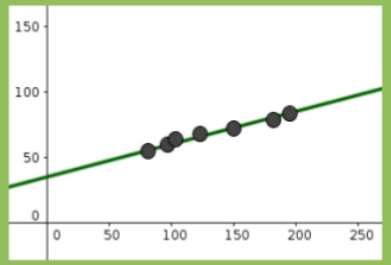
… en 1898, Amos Dolbear “tomó medidas” al respecto

Grillo moteado del arbusto;
via Johan J.Ingles-Le Nobel (Flickr)



Dr. Amos Dolbear; Tufts Digital Library

Crick	Temperatura
81	54.5
97	59.5
103	63.5
123	67.5
150	72.0
182	78.5
195	83.0



$$T_F = 50 + \left(\frac{N_{60} - 40}{4}\right)$$

N60: núm de criquidos (aleteos) por minuto
TF: temperatura (grados Fahrenheit)

El problema

Queremos relacionar dos variables cuantitativas:

1. Variable explicativa (independiente)
2. Variable respuesta (dependiente)

Construir **modelo**

Del tipo $y = f(x)$

Fuentes de ruido

- No todos los individios responden igual a un estímulo
- Variables no consideradas

Las relaciones $y = f(x)$

son unívocas pero obvian el ruido

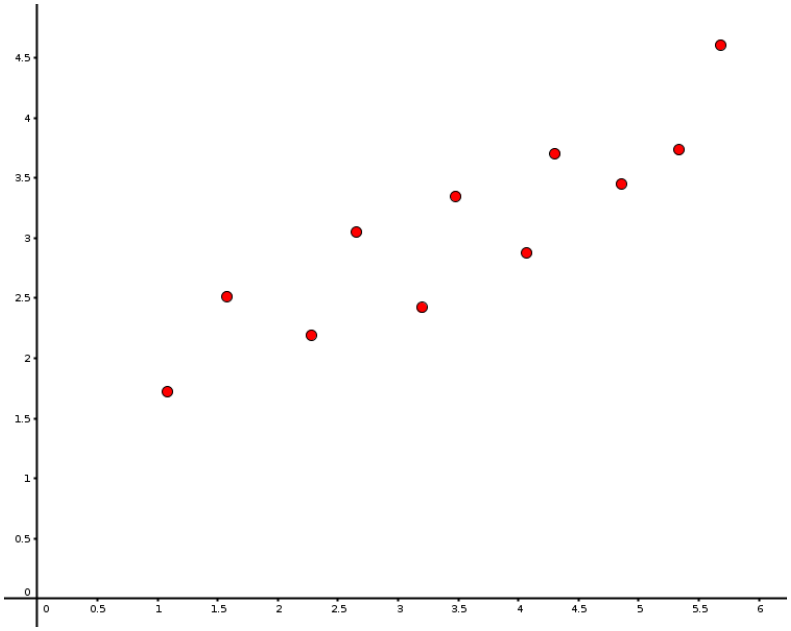
Empezaremos con **rectas**

La recta de regresión

Para estudiar la relación entre las variables X e Y disponemos de una muestra de pares de valores medidos simultaneamente

$$(x_1, y_1),\ (x_2, y_2), \cdots ,\ (x_n, y_n)$$

Análisis exploratorio: diagrama de dispersión (nube de puntos)

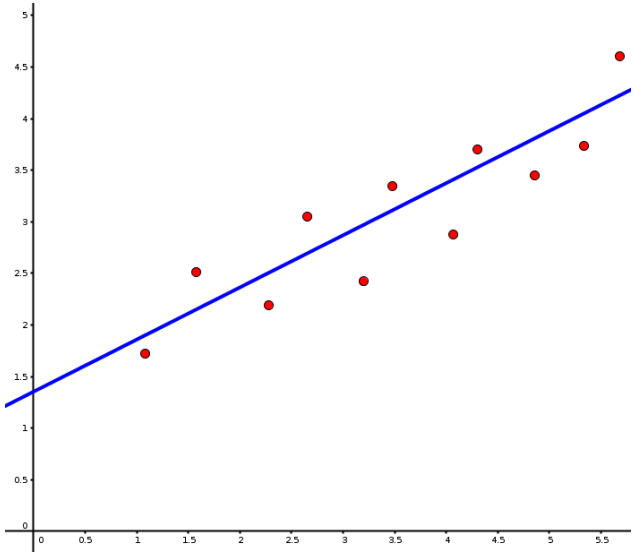


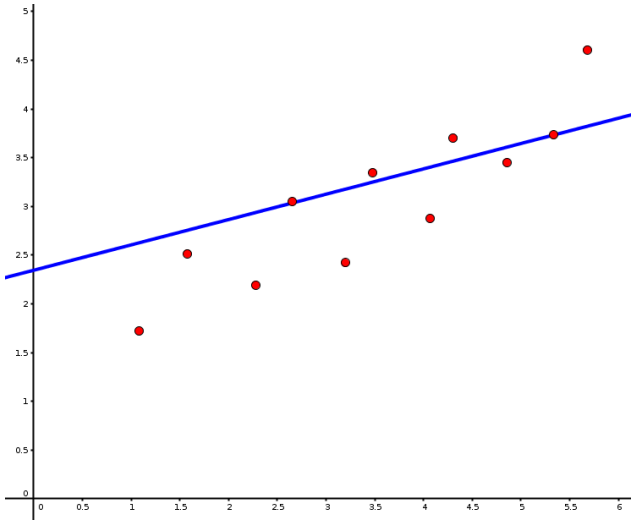
La recta de regresión

Ideas buscar una recta

$$y(x) = b_0 + b_1 \cdot x$$

que mejor “aproxime” (“describa”) la nube de puntos



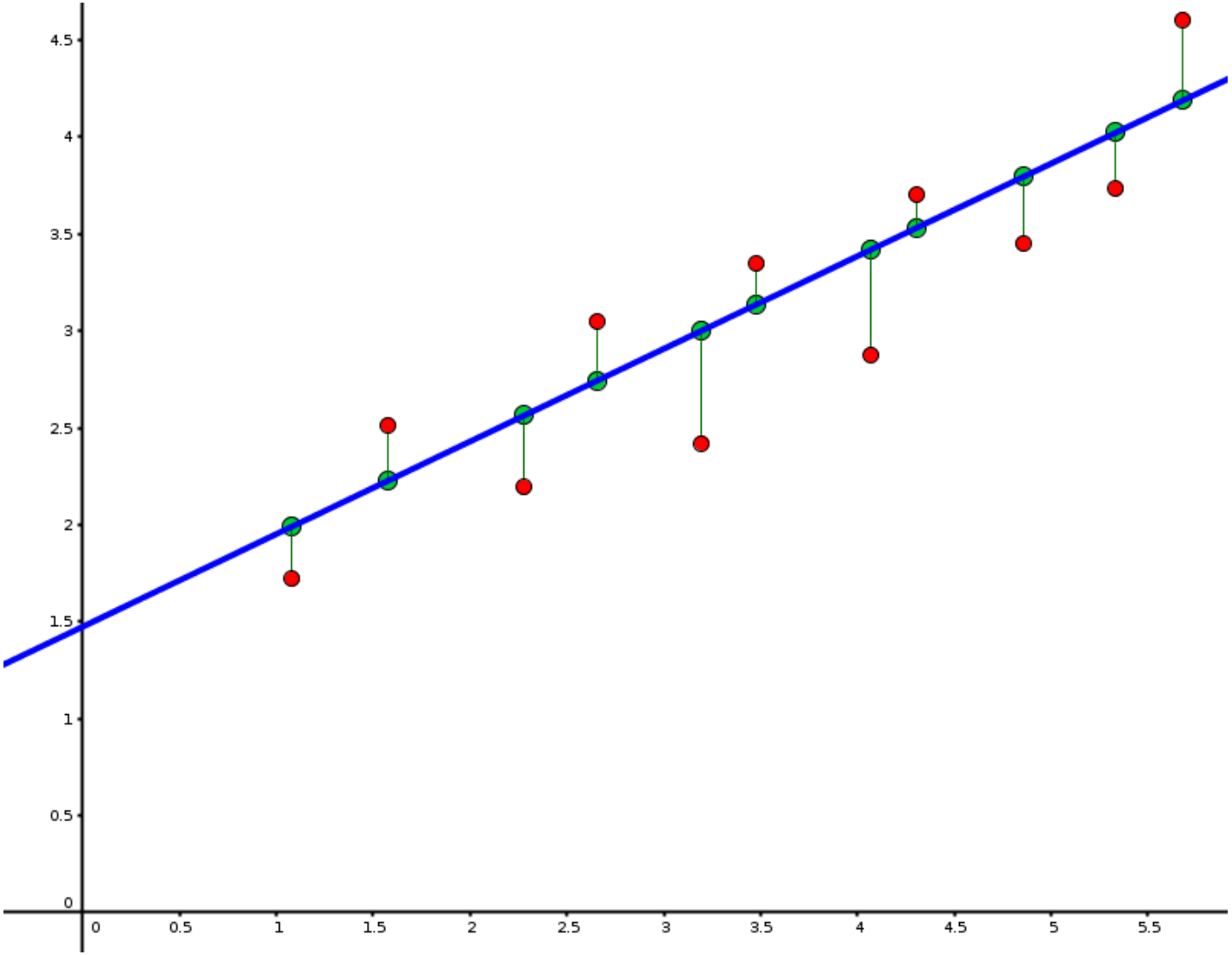


1. ¿Cómo elegir la recta?
2. ¿Qué información proporciona? ¿Cómo se usa?
3. ¿Cómo de buena es la descripción de los datos por la recta?

La recta de regresión

En rojo: valores observados **En verde**: valores predichos

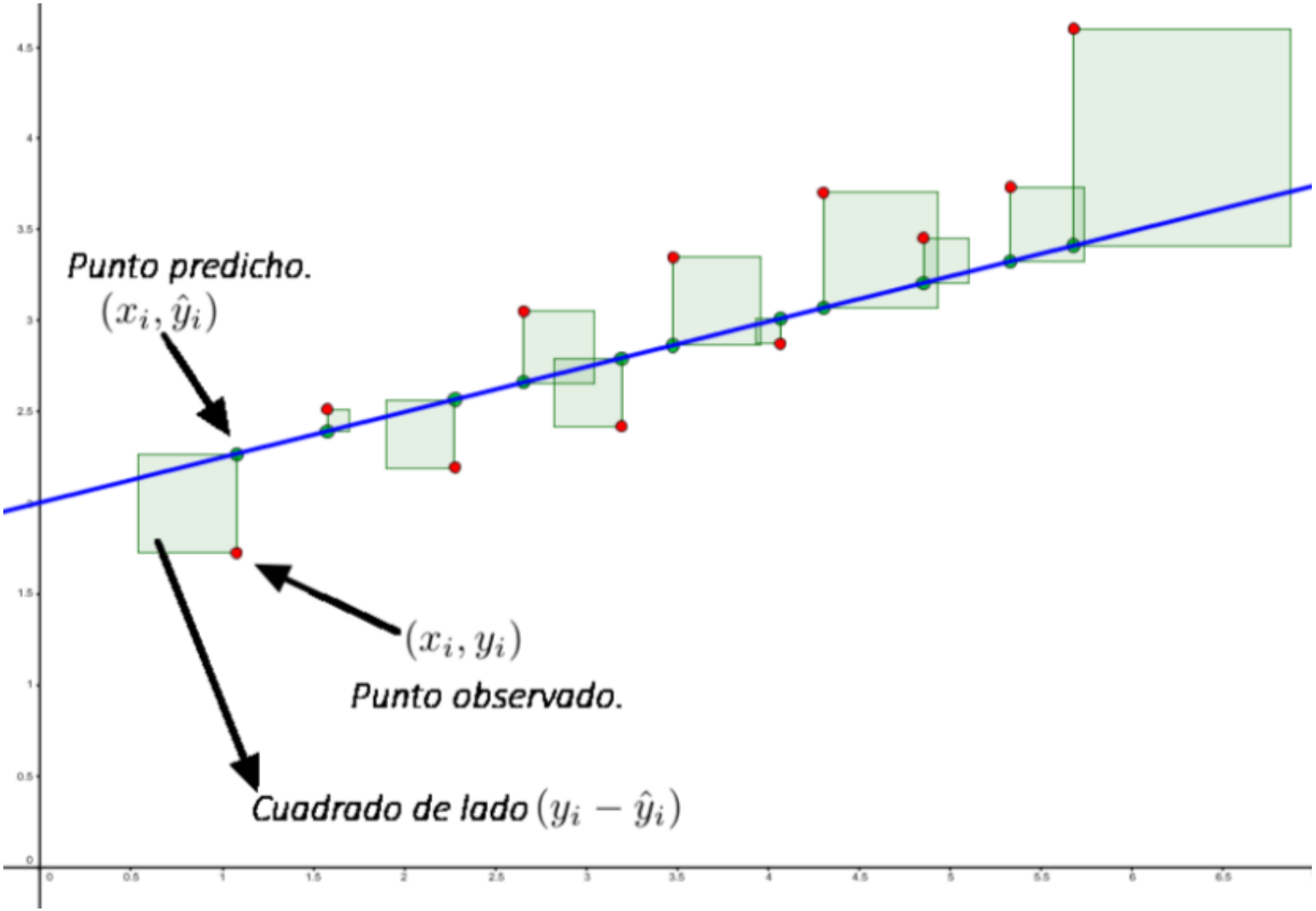
Residuo = observado - predicho



Minimizar los residuos (en cierto sentido)

La recta de regresión

Buscar b_0 y b_1 que minimizan la media de los residuos (errores) al cuadrado



Interpretación geométrica del EC y del ECM

Coeficientes de la recta de regresón
comparar los **valores observados** para la variable respuesta

$$y_1, y_2, \cdots, y_n$$

para cada valor de la variable explicativa con los **valores predichos** (o **ajustados**) por la recta

$$\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$$

donde

$$\hat{y}_j = b_0 + b_1 \cdot x_j, \qquad j = 1, 2, \cdots, n$$

y estudiamos la diferencia entre los valores observados y los valores predichos por el modelo, que llamaremos **residuos**

$$e_1 = y_1 - \hat{y}_1, \qquad e_2 = y_2 - \hat{y}_2, \qquad \cdots, \qquad e_n = y_n - \hat{y}_n$$

Cada elección de b_0 y b_1 produce residuos diferentes

Error cuadrático

Dado el conjunto de puntos

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n),$$

si consideramos los valores predichos

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n,$$

con

$$\hat{y}_j = b_0 + b_1 \cdot x_j, \qquad \text{para } j = 1, \ldots, n,$$

entonces el **error cuadrático** (*sum of squared errors*) de la recta $y = b_0 + b_1 \cdot x$ es:

$$\mathrm{EC}(y = b_0 + b_1 \cdot x) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2.$$

El error cuadrático medio ECM es el promedio muestral:

$$ECM = \frac{EC}{n-1}$$

Recta de regresión (o de mínimos cuadrados). Covarianza

Dado el conjunto de puntos $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$, la **recta de regresión o de mínimos cuadrados** (*regression line* o *line of best fit*) es la recta que minimiza el ECM y está dada por

$$(y - \bar{y}) = \frac{Cov(x, y)}{s^2(x)} \cdot (x - \bar{x}),$$

siendo

$$Cov(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

es la **covarianza muestral** (*covariance*) de $(x_1, y_1), \ldots, (x_n, y_n)$.

Si la recta es $y = b_0 + b_1 \cdot x$, entonces

$$b_1 = \frac{Cov(x, y)}{s^2(x)}, \qquad b_0 = \bar{y} - \frac{Cov(x, y)}{s^2(x)} \cdot \bar{x}.$$

Ejemplo: disponemos de los datos

Núm cricks/minuto	81	97	103	123	150	182	195
Temperatura (F)	54.5	59.5	63.5	67.5	72.0	78.5	83.0

partir de estos datos calculamos

$$\bar{x} \approx 133, \quad s^2(x) \approx 1922.333$$

$$\bar{y} \approx 68.36, \quad s^2(y) \approx 104.23,$$

$$Cov(x, y) \approx 443.421$$

Por lo tanto,

$$b_1 \approx -0.4896, \qquad b_0 \approx 2.766,$$

y la recta de regresión es

$$Temperatura = 37{,}7.766 + 0.23 \cdot Cricks$$

¿Intrepretación de b_1?

¿Cuánto aumenta la temperatura al aumentar los grillos en una unidad el número de cricks/minuto

Observaciones: uso de la recta de regresión

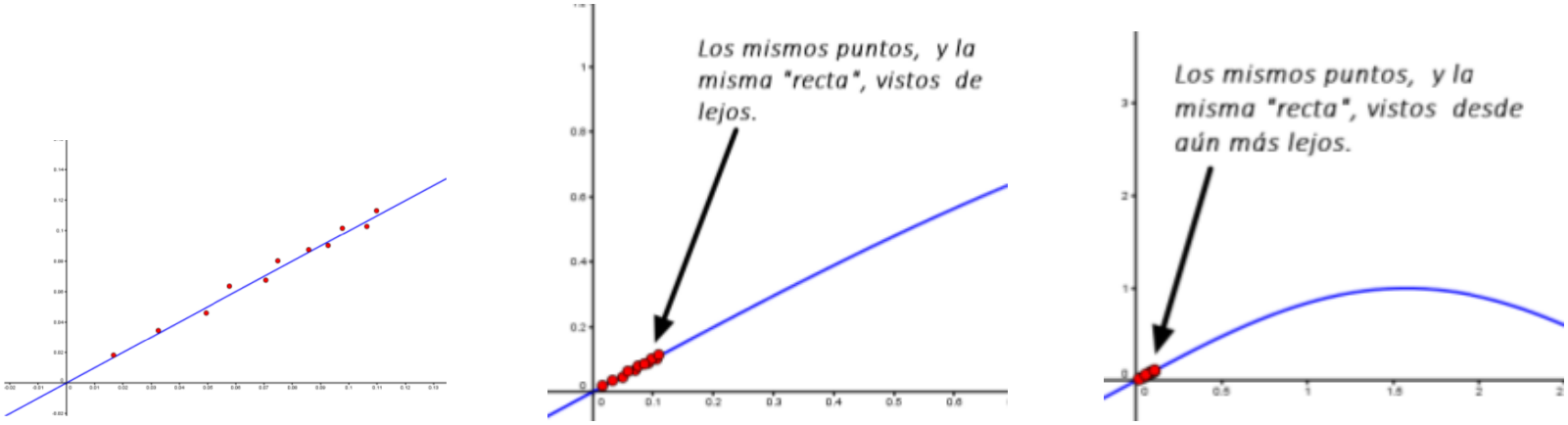
- Podemos usar la recta para estimar el valor de la variable respuesta y para valores no observados de la variable explicativa x.

- Sabemos que las estimaciones puntuales son, por naturaleza, imprecisas. Mejor usar un intervalo

Nunca, bajo ningún concepto, está justificado el uso de la recta para predecir valores de y correspondientes a valores de x fuera del intervalo (m_x, M_x). Hacer eso se denomina **extrapolación** y se considera uno de los errores más graves que pueden cometer en el contexto del uso de la recta de regresión.

Observaciones: ¿por qué usar rectas?

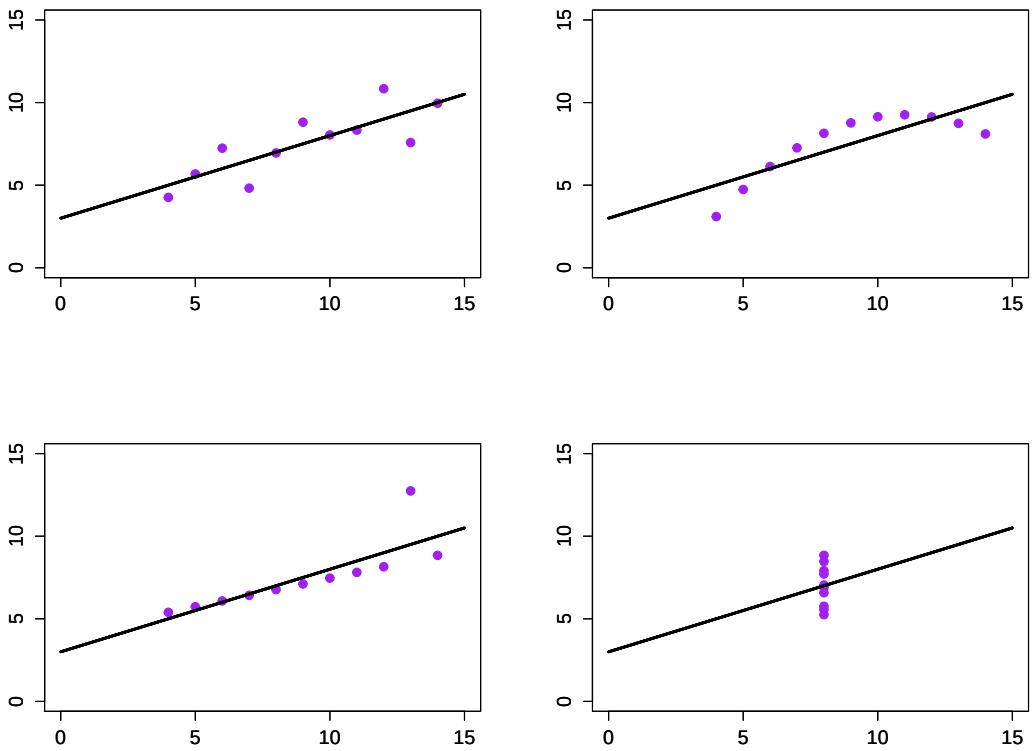
Frente a curvas más complejas, recuerda que la recta tangente es una buena **aproximación local**



Zooms de una función en un entorno de $x = 0$

Observaciones ¿siempre tiene sentido?

Para cualquier conjunto de datos con $s^2(X) \neq 0$ se puede calcular la recta de regresión; ¿es siempre útil?



Las 4 nubes de puntos tienen asociada la misma recta de regresión!!
Asignar una medida a la **bondad del ajuste**