2do C 2021

Laboratorio 1 (Físicos)

REGRESIÓN LINEAL

Cátedra Goyanes

Repasemos…

- Definimos un sistema físico a partir de parámetros y variables.

  Medimos de manera sistemática
- Expresamos resultados como un intervalo → $x = (x_0 \pm \Delta x)$
- Utilizamos herramientas e indicadores estadísticos
- El error (o incerteza) absoluta $\boxed{\Delta x = \sqrt{\sigma_N^2 + \sigma_e^2}}$
- Mediciones indirectas → $y = f(x_i)$
- Propagación de errores: $\Delta y^2 = \sum \left(\frac{\partial y}{\partial x_i} \Delta x_i \right)^2$

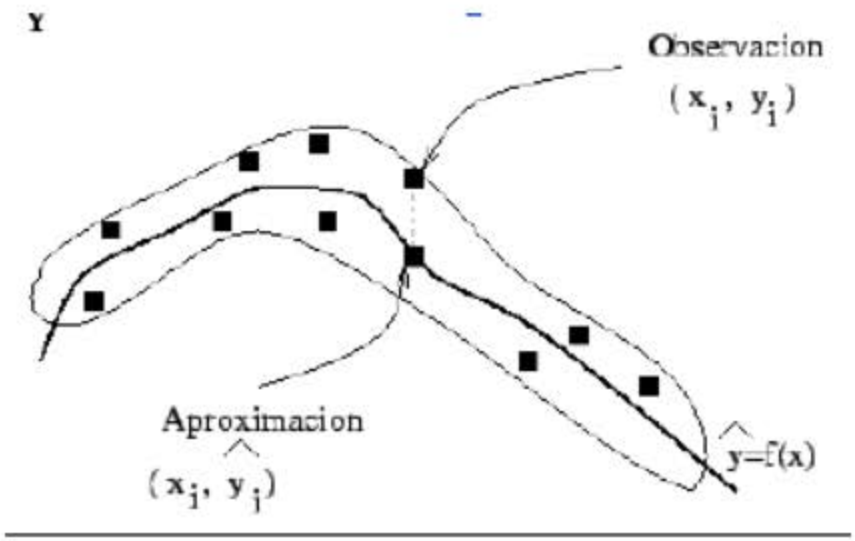
  (variables independientes)

Modelo de Regresión

El objetivo de un **modelo de regresión** es tratar de explicar la relación que existe entre una variable dependiente (Y) y un conjunto de variables independientes (X1,..., Xn).

En un **modelo de regresión simple** hay 1 sola variable independiente X_i

Mediante las técnicas de regresión de una variable Y sobre una variable X, buscamos una función que sea una buena aproximación de una nube de puntos (x_i, y_i), mediante una curva del tipo:



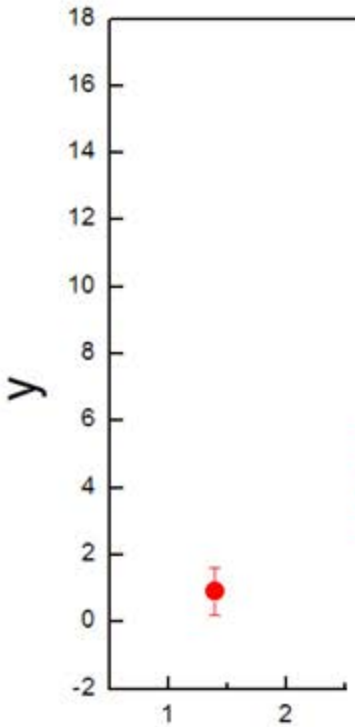
$$\hat{Y} = f(X)$$

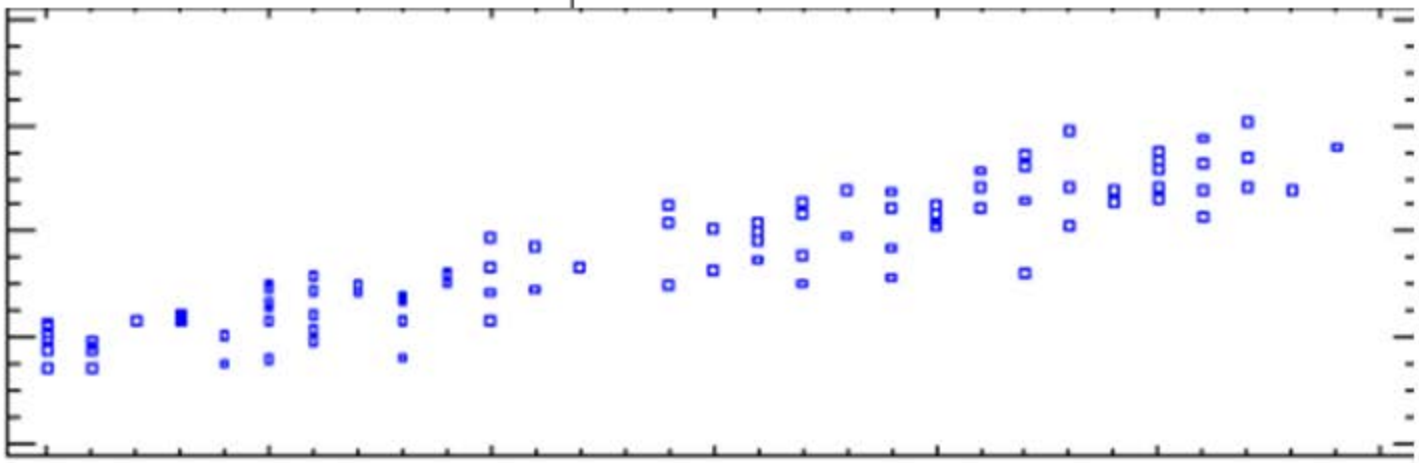
Regresión Lineal (x_i, y_i)

No determinista: Conocido el valor de X, el valor de Y no queda perfectamente establecido. Son del tipo:

$$y = f(x) + u$$

donde u es una perturbación desconocida (variable aleatoria).



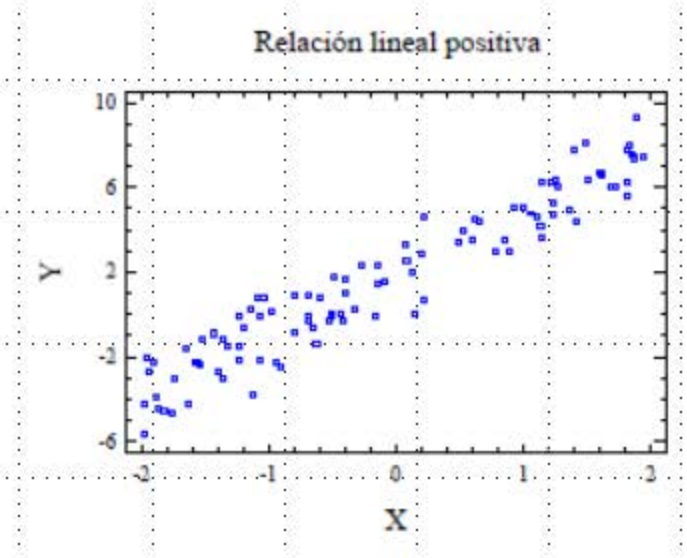


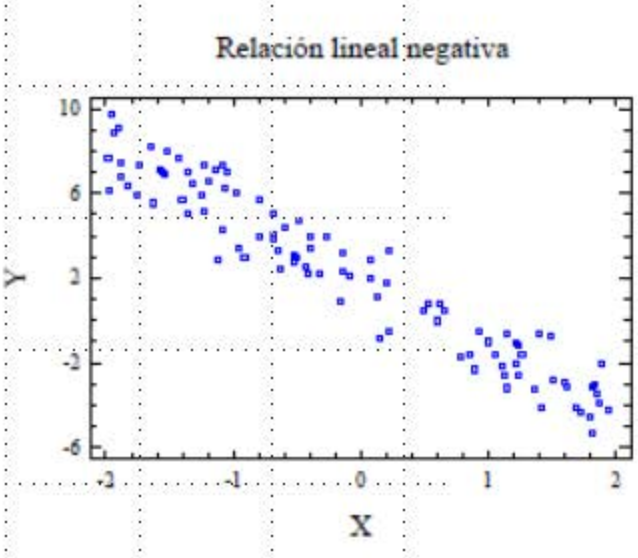
Tipos de relación

- Lineal: Cuando la función $f(x)$ es lineal,

$$f(x) = \beta_0 + \beta_1 x$$

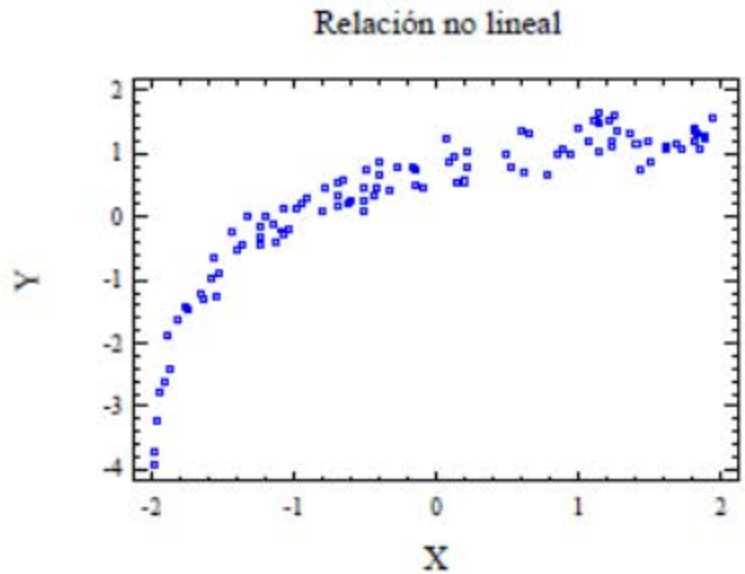
  - Si $\beta_1 > 0$ hay relación lineal positiva.
  - Si $\beta_1 < 0$ hay relación lineal negativa.





Tipos de relación

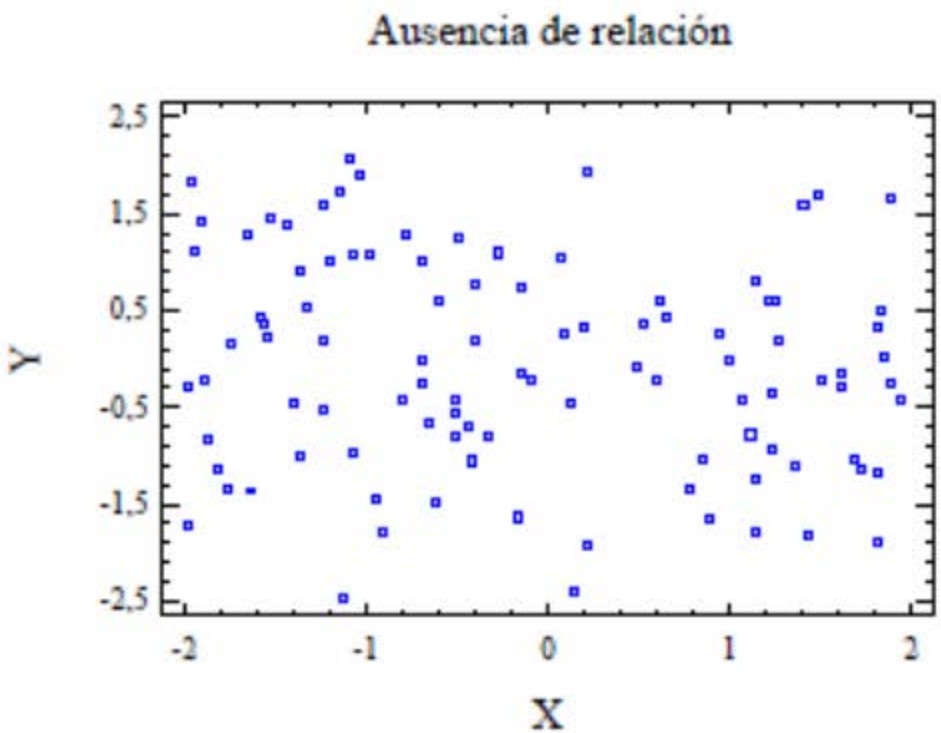
- No lineal: Cuando la función $f(x)$ no es lineal. Por ejemplo, $f(x) = log(x)$, $f(x) = x^2 + 3, \ldots$



Los datos no tienen un aspecto recto.

Tipos de relación

- Ausencia de relación: Cuando $f(x) = 0$.



La covarianza

Una medida de la dependencia lineal es la covarianza:

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

- Si hay relación lineal positiva, la covarianza será positiva y grande.
- Si hay relación lineal negativa, la covarianza será negativa y grande en valor absoluto.
- Si hay no hay relación entre las variables o la relación es marcadamente no lineal, la covarianza será próxima a cero.

PERO la covarianza depende de las unidades de medida de las variables.

El coeficiente de correlación lineal

Una medida de la dependencia lineal que no depende de las unidades de medida es el coeficiente de correlación lineal:

$$r_{(x,y)} = cor\,(x,y) = \frac{cov\,(x,y)}{s_x s_y}$$

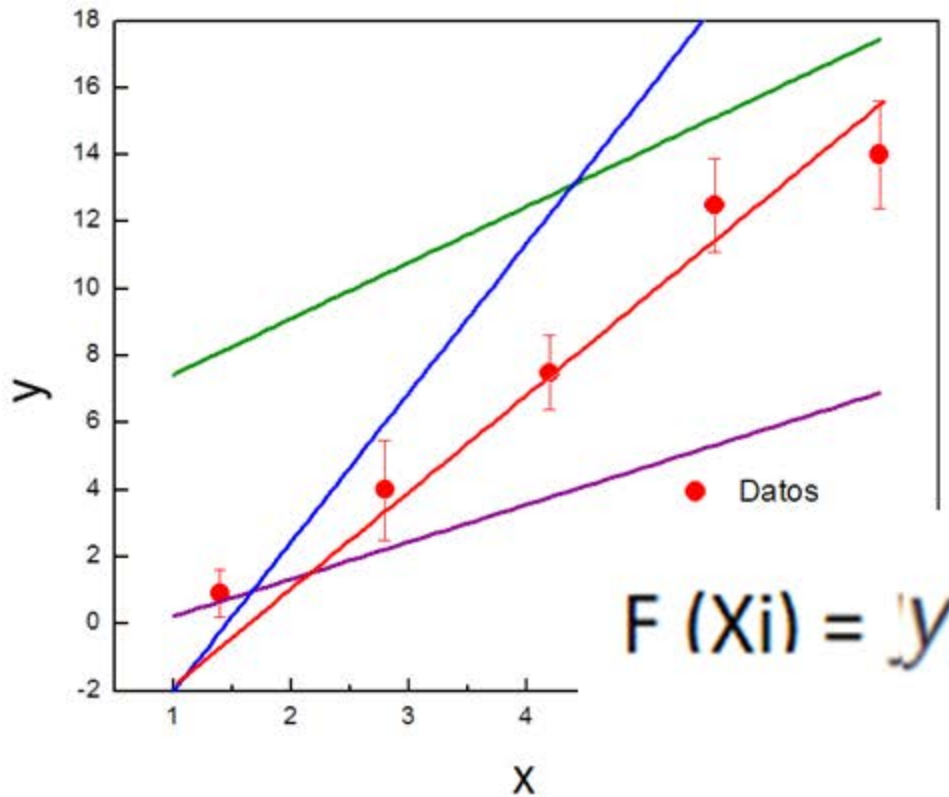
donde:

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

- $-1 \leq cor\,(x,y) \leq 1$
- $cor\,(x,y) = cor\,(y,x)$

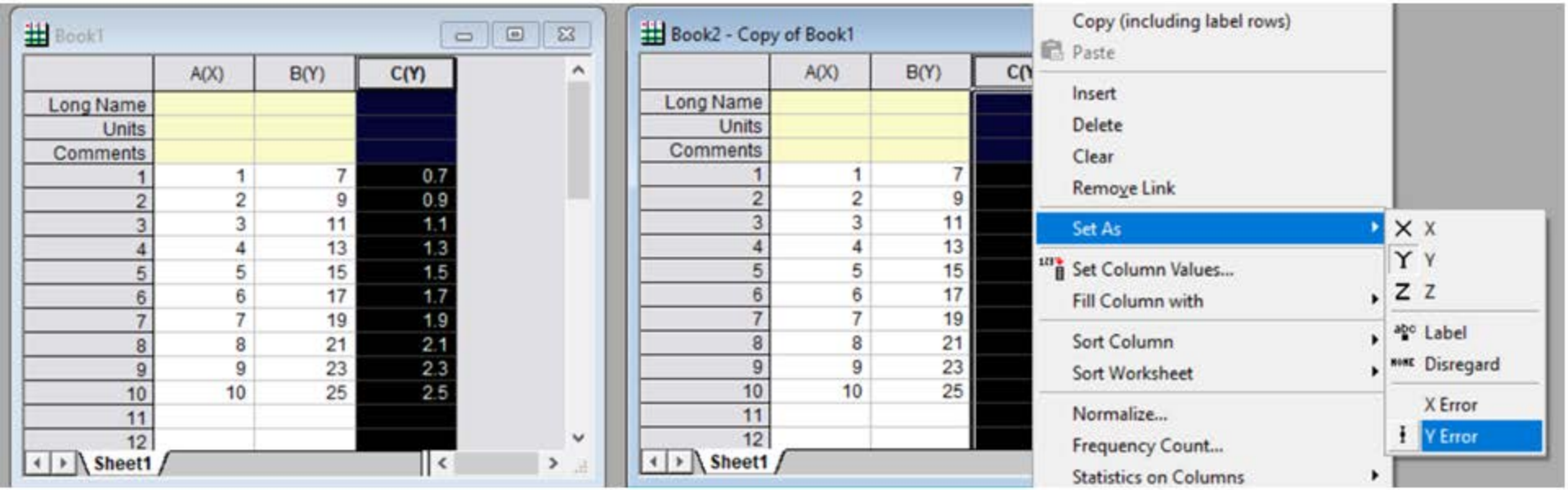
Regresión Lineal (x_i, y_i)

Para hacer una estimación del modelo de regresión lineal simple, trataremos de buscar una recta de la forma



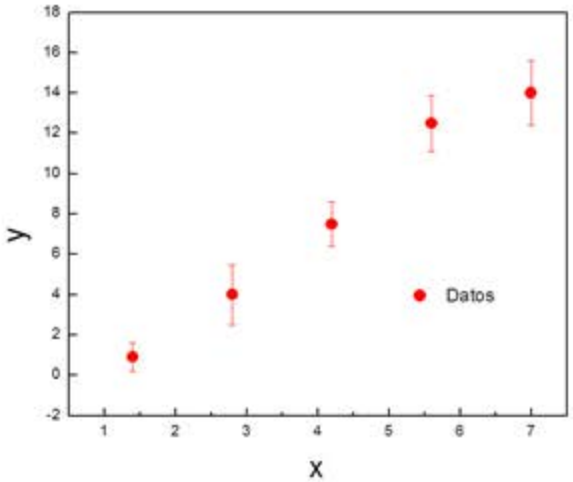
$$F(X_i) = y_i = \beta_0 + \beta_1 x_i + u_i$$

¿Cómo se grafica? ¿Cómo le digo que columna es el error?



¿ Qué error tengo que graficar? El ABSOLUTO

El modelo de regresión lineal simple



El modelo de regresión lineal simple supone que,

$$\boxed{y_i = \beta_0 + \beta_1 x_i + u_i}$$

donde:

- y_i representa el valor de la variable respuesta para la observación i-ésima.
- x_i representa el valor de la variable explicativa para la observación i-ésima.
- u_i representa el error para la observación i-ésima que se asume normal,

$$u_i \sim N(0, \sigma)$$

- β_0 y β_1 son los coeficientes de regresión:
  - β_0 : intercepto
  - β_1 : pendiente

Los parámetros que hay que estimar son: β_0, β_1 y σ.

Hipótesis del modelo de regresión lineal simple

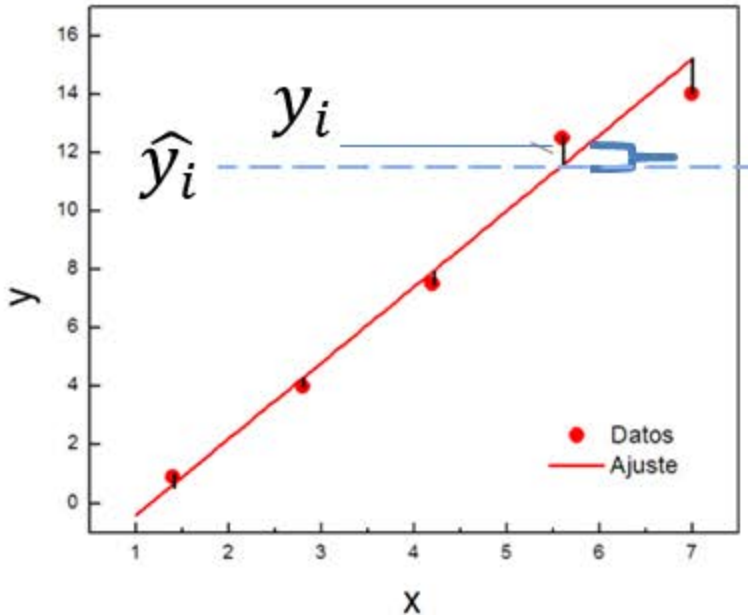
Independencia

- Los datos deben ser independientes.
- Una observación no debe dar información sobre las demás.

Normalidad

- Se asume que los datos son normales a priori.

$$\boxed{y_i = a\,x_i + b + e_i}$$



$$a\,x_i + b = \widehat{y_i}$$

$e_i \sim N(0, \sigma^2)$ "Residuos"

$$\boxed{e_i \equiv y_i - \widehat{y_i}}$$

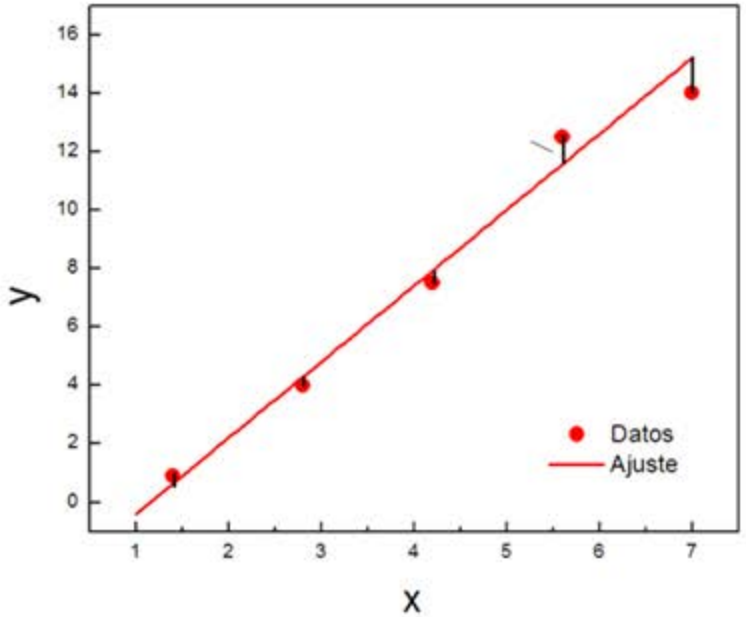
Estimadores de mínimos cuadrados

Gauss propuso en 1809 el método de mínimos cuadrados para obtener los valores $\boldsymbol{a}$ y **b** que mejor se ajustan a los datos:

$$\widehat{y_i} = a\, x_i + b$$

El método consiste en minimizar la suma de los residuos al cuadrado (S^2):

$$S^2 = \sum_i^N e_i^2 = \sum_i^N (y_i - \widehat{y_i})^2 = \sum_i^N [y_i - (a\, x_i + b)\,]^2$$

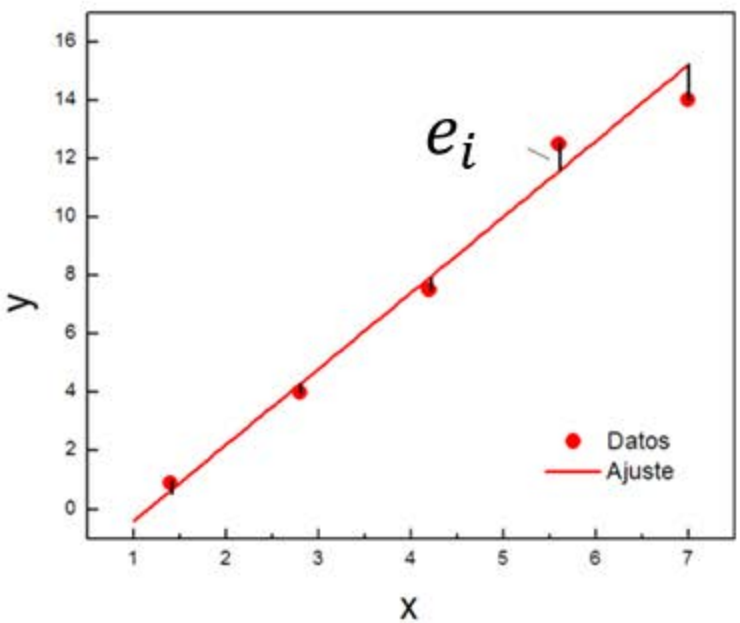


$$a\, x_i + b = \widehat{y_i}$$

"Residuos"

$$\boxed{e_i \equiv y_i - \widehat{y_i}}$$

Es decir, minimizar la suma de los cuadrados de las diferencias entre los valores "experimentales" observados (y_i) y los valores $(\hat{y}_i)$ nados



$$S^2 = \sum_i^N e_i^2 = \sum_i^N [y_i - (a.x_i + b)\,]^2$$

Debo minimizar $S^2(a, b)$

Incógnitas: a y b

¿Cómo cálculo a y b en el ajuste?
Minimizo S^2

$$\left.\begin{aligned} \frac{\partial S^2}{\partial a}(a,b) &= 0 \\ \frac{\partial S^2}{\partial b}(a,b) &= 0 \end{aligned}\right\} \xrightarrow{\cdots}$$

$$\boxed{\begin{aligned} a &= \frac{N\sum(x_i\, y_i) - (\sum x_i).(\sum y_i)}{N\sum {x_i}^2 - (\sum x_i)^2} \\ b &= \frac{(\sum {x_i}^2).(\sum y_i) - (\sum x_i).(\sum x_i y_i)}{N\sum {x_i}^2 - (\sum x_i)^2} \end{aligned}}$$

¿Cuáles son los errores de los parámetros obtenidos a y b?

Propagamos errores:

$$\sigma_a^2 = \sum \left(\frac{\partial a}{\partial y_i} \sigma_y \right)^2 \qquad \sigma_b^2 = \sum \left(\frac{\partial b}{\partial y_i} \sigma_y \right)^2$$

Si $\sigma_{y_1} = \sigma_{y_2} = \ldots\ldots = \sigma_{yN} = \sigma_y$

$$\boxed{\sigma_A = \sigma_y \sqrt{\frac{\sum x^2}{\Delta}}} \qquad \boxed{\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}}$$

Donde: $\Delta = N \sum x^2 - (\sum x)^2$

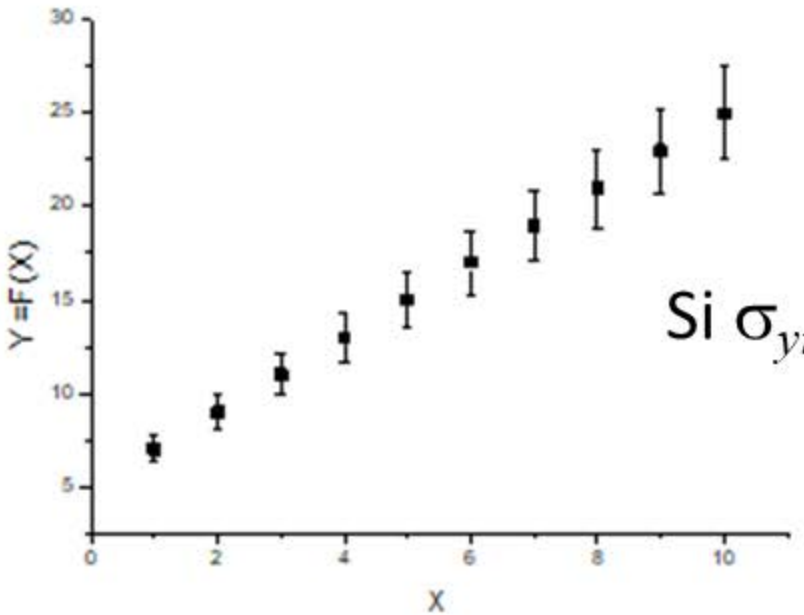
Estimación del σ_y

- Asumimos que todos lo y_i tienen la misma función distribución normal con el mismo σ_y

- La dispersión de los datos alrededor de la recta estimada tendrá la misma desviación estándar

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - A - Bx_i)^2},$$

Hasta acá no tuvimos en cuenta el error de los datos. Algunos datos podrían ser más confiables que otros, y queremos que "pesen" más.



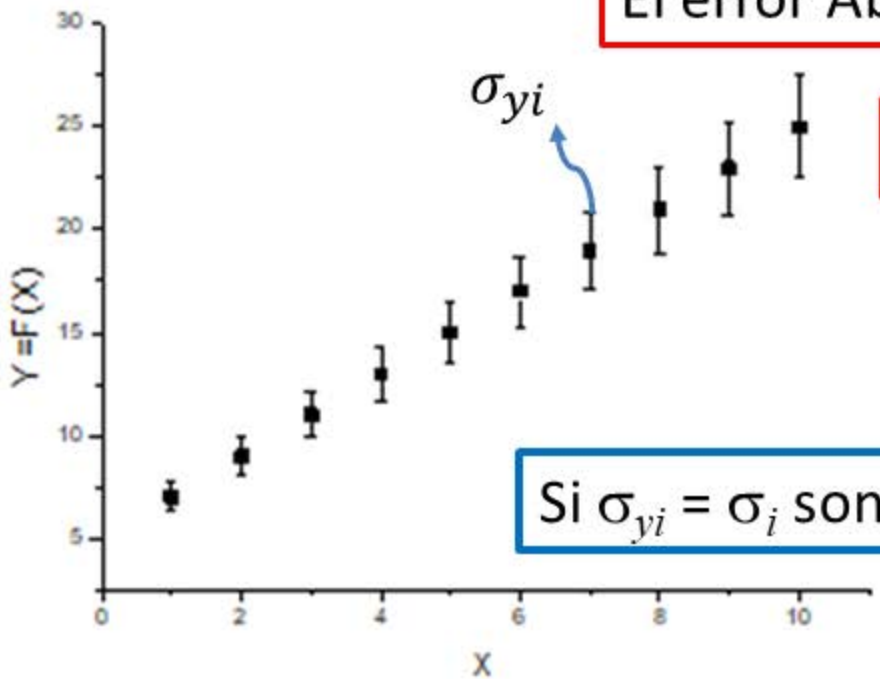
Si $\sigma_{yi} = \sigma_i$ son distintos:

Peso

$$w_i = \frac{1}{\sigma_i^2}$$

Ajuste de cuadrados mínimos pesado con la incertidumbre

El error Absoluto cambia de dato a dato



¿Cómo tengo en cuenta eso?

CM PONDERADOS

Si $\sigma_{yi} = \sigma_i$ son distintos

$$w_i = \frac{1}{\sigma_i^2}$$

CM sin ponderar

$$S^2 = \sum_i^N e_i^2 = \sum_i^N [y_i - f(x_i)]^2$$

CM PONDERADOS

$$S^2 = \sum_i^N \left[\frac{y_i - f(x_i)}{\sigma_{yi}}\right]^2 = \sum_i^N \frac{[y_i - f(x_i)]^2}{\sigma_{yi}^2}$$

$$S^2 = \sum_i^N \left[\frac{y_i - f(x_i)}{\sigma_{yi}}\right]^2$$

$$= \sum_i^N \frac{[y_i - f(x_i)]^2}{{\sigma_{yi}}^2}$$

$$a = \frac{1}{\Delta}\left[\sum \frac{1}{{\sigma_{yi}}^2} \sum \frac{x_i y_i}{{\sigma_{yi}}^2} - \sum \frac{x_i}{{\sigma_{yi}}^2} \sum \frac{y_i}{{\sigma_{yi}}^2}\right]$$

$$b = \frac{1}{\Delta}\left[\sum \frac{{x_i}^2}{{\sigma_{yi}}^2} \sum \frac{y_i}{{\sigma_{yi}}^2} - \sum \frac{x_i}{{\sigma_{yi}}^2} \sum \frac{x_i y_i}{{\sigma_{yi}}^2}\right]$$

$$\Delta = \sum \frac{1}{{\sigma_{yi}}^2} \sum \frac{{x_i}^2}{{\sigma_{yi}}^2} - \left(\sum \frac{x_i}{{\sigma_{yi}}^2}\right)^2$$

Cuadrados mínimos solo considera errores en eje Y

- Se debe cumplir que:

$$\boxed{\Delta x \frac{dy}{dx} \ll \Delta y}$$

En caso contrario, debemos **invertir los ejes**.

- Si los dos términos son del **mismo orden**, podemos "trasladar" el error en x al error en y:

$$\Delta y_{\text{nuevo}}^2 = \left(\Delta x \frac{dy}{dx}\right)^2 + \Delta y_{original}^2$$

- Si aproximamos por una recta y **pasa por el origen**,

entonces el criterio se puede expresar como: $\varepsilon_x \ll \varepsilon_y$

Ajuste de cuadrados mínimos pesado con la incertidumbre

Si $\sigma_{yi} = \sigma_i$ son distintos

$$w_i = \frac{1}{\sigma_i^2}$$

$$\Delta = \sum w \sum wx^2 - (\sum wx)^2$$

$$\Delta = N\sum x^2 - (\sum x)^2$$

$$\sigma_A = \sigma_y \sqrt{\frac{\sum x^2}{\Delta}}$$

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}$$

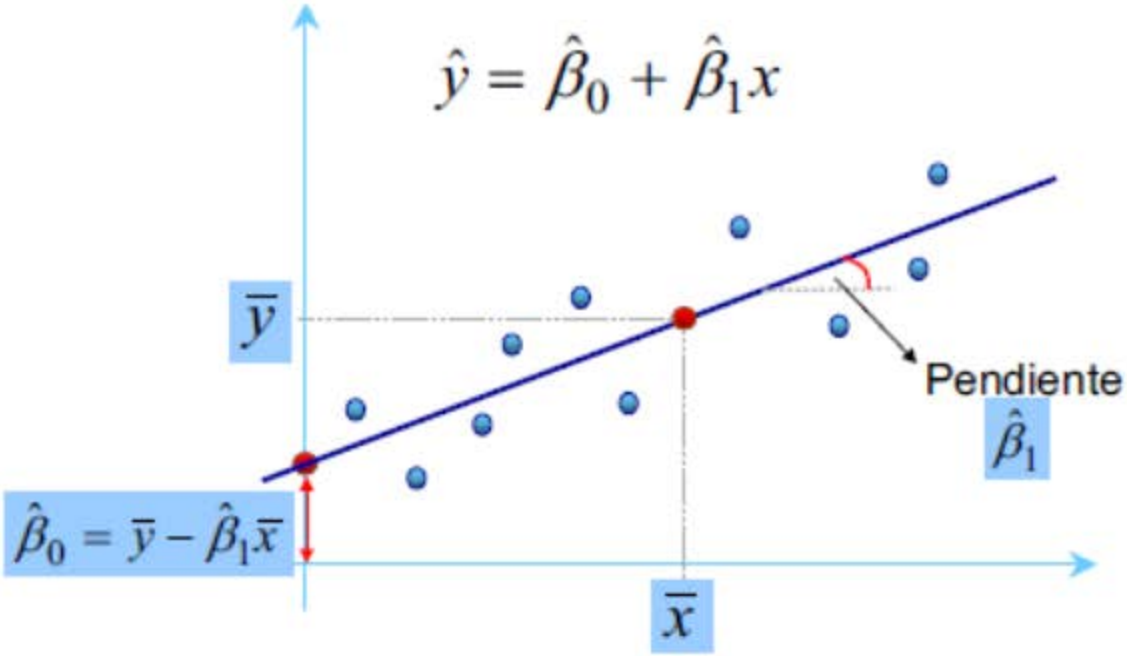
$$\sigma_A = \sigma_y \sqrt{\frac{\sum wx^2}{\Delta}}$$

$$\sigma_B = \sigma_y \sqrt{\frac{\sum w}{\Delta}}$$

Estimadores de mínimos cuadrados

$$\hat{\beta}_1 = \frac{cov(x,y)}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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



Covarianza

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \quad S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}, \quad S_Y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n}, \quad S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n}.$$

Observación: La recta de regresión pasa siempre por el centro de gravedad de la nube de puntos, es decir por el punto . $(\overline{X}, \overline{Y})$

S_{XY} es la **covarianza muestral** entre *X* e *Y*.

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

La covarianza es un valor que indica el grado de interralación entre dos variables respecto de sus propias medias o promedios.

Dicho de otra forma, la covarianza, como la misma expresión lo indica, nos dice si dos variables tienen que ver, de alguna manera, la una con la otra

S_X^2 es la **varianza muestral** de *X*

Coeficiente de Correlación Lineal r

El coeficiente de correlación lineal entre *X* e *Y* viene dado por

$$r = \frac{S_{XY}}{S_X S_Y},$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

Mide la dependencia lineal que existe entre las dos variables.

Propiedades del coeficiente de correlación:

a) No tiene dimensión, y siempre toma valores en [-1,1].
b) Si las variables son independientes, entonces r=0, pero el inverso no tiene por qué ser cierto.
c) Si existe una relación lineal exacta entre *X* e *Y*, entonces *r* valdría 1 (relación directa) ó -1 (relación inversa).
d) Si $r > 0$, esto indica una relación directa entre las variables (es decir, que si aumentamos *X*, también aumenta *Y*).
e) Si $r < 0$, la correlación entre las variables es inversa (si aumentamos una, la otra disminuye).

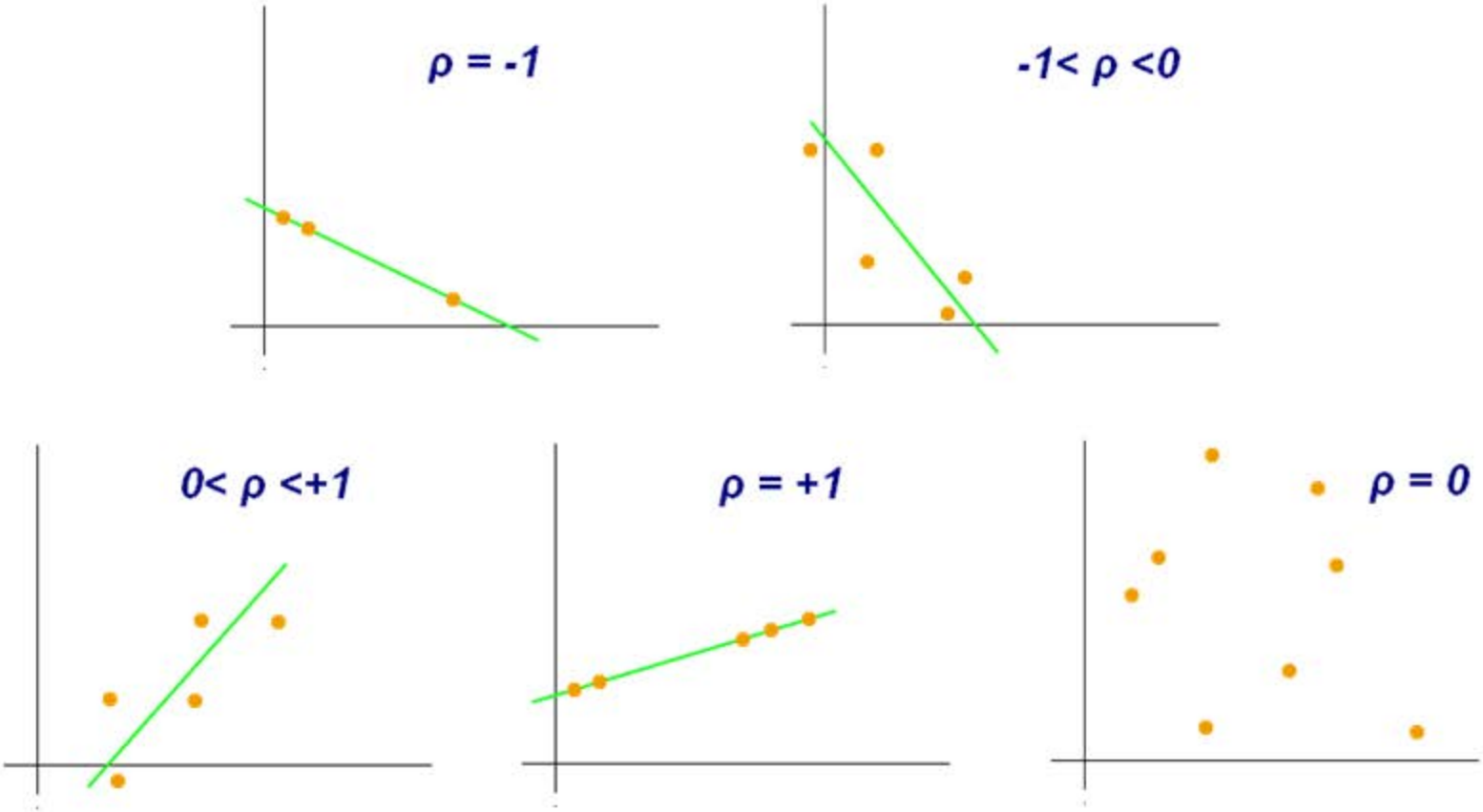
Coeficiente de Correlación de Pearson ρ

Mide el grado de correlación lineal entre dos variables

$$-1 < \rho < 1$$

$\rho = |1|$ indica que los datos están perfectamente alineados

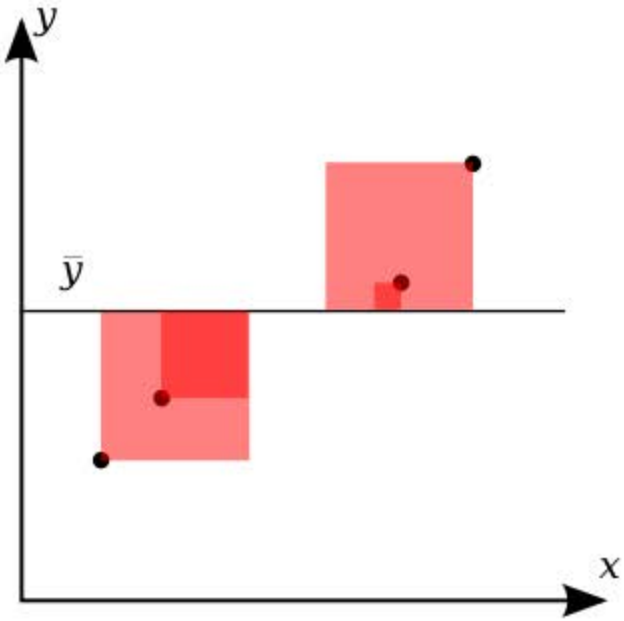
$\rho = 0$ indica que no hay correlación lineal entre las variables

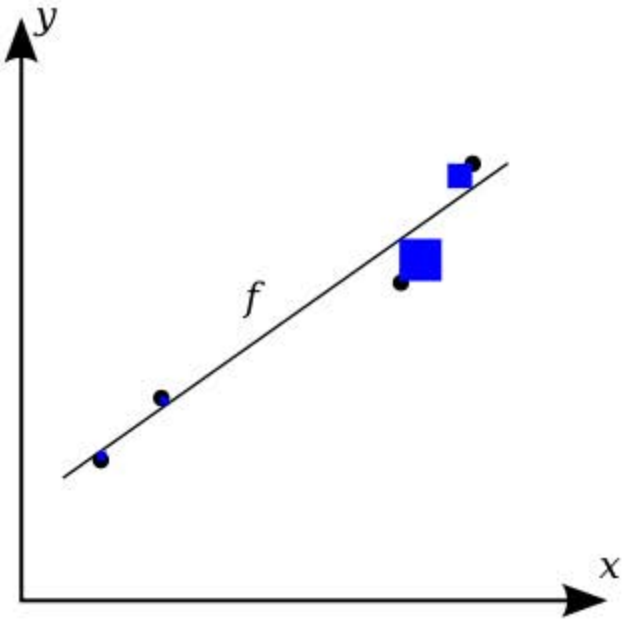


Coeficiente de Determinación R^2 (r-squared)

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Mientras más cercano a 1 mejor es el ajuste ($R^2 = r^2$)





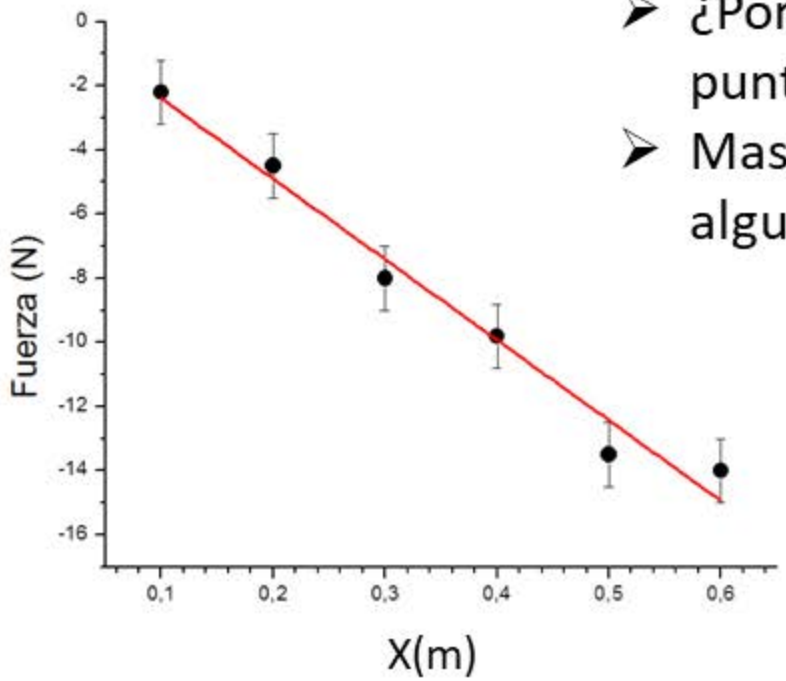
$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2 \qquad SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

Para cuantificar la bondad del ajuste de un modelo, lineal o no, se utiliza el coeficiente de determinación lineal R_2

https://www.youtube.com/watch?v=MtDXpWC_RLc

PARA PENSAR y CONTESTAR LA PROXIMA

Supongamos que quiero sacar la constante elástica de un resorte



- ¿Tiene sentido hacer tantas mediciones?
- ¿Por qué no la saco directamente de un punto?
- Mas allá de determinar el rango lineal, ¿tiene alguna otra ventaja?

- En el caso, ¿estaría bien sacar K como el cociente?
- ¿Qué pasa con los errores a medida que aumento el numero de mediciones?

¿Y si el modelo no es lineal?

Muchas veces puede linealizarse la ecuación

$$y = a.x^3$$

Se grafica $y\ vs\ x^3$

$$\Delta({x_i}^3) = \sqrt{\left(\frac{\partial(x^3)}{\partial x}\Delta x_i\right)^2}$$

$$y = a.b^x$$

$$\ln(y) = \ln(a.b^x)$$

$$\ln(y) = \ln(a) + x.ln(b)$$

Se grafica $\ln(y)\ vs\ x$

Δ (a) = ?????

Δ (b) = ?????

¿Qué se obtiene de CM????

$$y(x) = a.x^b$$

$$\ln(y) = \ln\left(a.x^b\right)$$

$$\ln(y) = \ln(a) + b.\ln(x)$$

Se grafica $\ln(y)$ vs $\ln(x)$

Justificación del principio de cuadrados mínimos

$$Prob_{A,B}(y_i) \propto \frac{1}{\sigma_y} e^{-(y_i - A - Bx_i)^2/2\sigma_y^2}$$

$$Prob_{A,B}(y_1, \ldots, y_N) = Prob_{A,B}(y_1) \cdots Prob_{A,B}(y_N)$$

$$\propto \frac{1}{\sigma_y^N} e^{-\chi^2/2},$$

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}.$$

El máximo de probabilidad corresponde al mínimo de χ^2