Error en ambas variables: Ajuste por regresión de distancia ortogonal (ODR)

Grupo 4 - Florencia Mastrandrea, Luis Diaz , Dante Brutti

Laboratorio 4A
Primer cuatrimestre 2024

Introducción

Introducción

ODR es un método de regresión que extiende la idea de cuadrados mínimos para contemplar errores en ambas variables.

Variables observadas $(X_i; Y_i) \quad i = 1, \ldots, n$

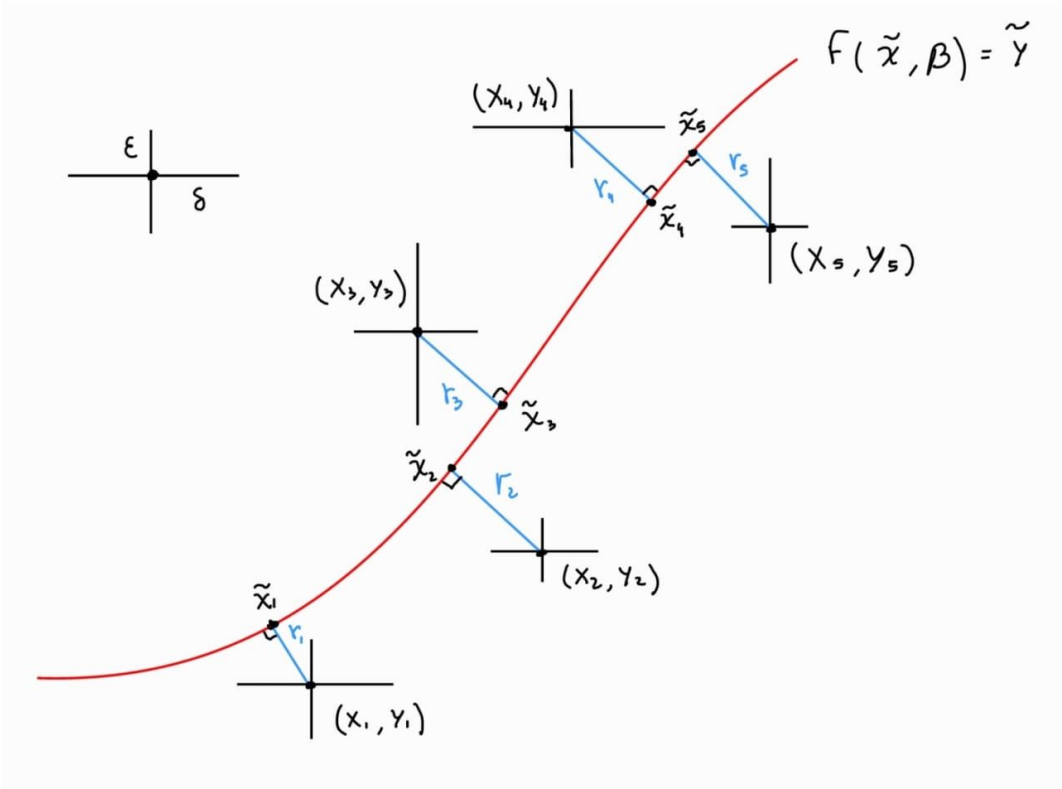
Variables verdaderas $(x_i; y_i) \quad i = 1, \ldots, n$

$$x_i = X_i + \delta_i$$
$$y_i = Y_i + \epsilon_i$$

$y_i = f(x_i; \beta)$ Se asume una dependencia entre los valores teóricos de x e y donde β son los parámetros de ajuste.

$Y_i = f(X_i + \delta_i; \beta) - \epsilon_i$ Se aplican las variables observadas al modelo teórico.

Para este modelo tanto la variable dependiente como los parámetros pueden asumirse no lineales.



Se llama “ortogonal” porque las distancias son perpendiculares a la tangente de la función.

Introducción

Se busca minimizar la distancia ortogonal total de los datos a la función dada.

$$r_i^2 = \min_{\epsilon_i, \delta_i} \sum_{i=1}^{n} \left\{ \epsilon_i^2 + \delta_i^2 \right\} \qquad \text{sujeto a} \qquad Y_i = f\left(X_i + \delta_i; \beta\right) - \epsilon_i$$

$$\Rightarrow \quad \min_{\beta_i, \delta_i} \sum_{i=1}^{n} \left\{ \left[f\left(X_i + \delta_i; \beta\right) - Y_i \right]^2 + \delta_i^2 \right\}$$

A esto se le agrega la consideración por los pesos de ambos errores, dados por la desviación estándar.

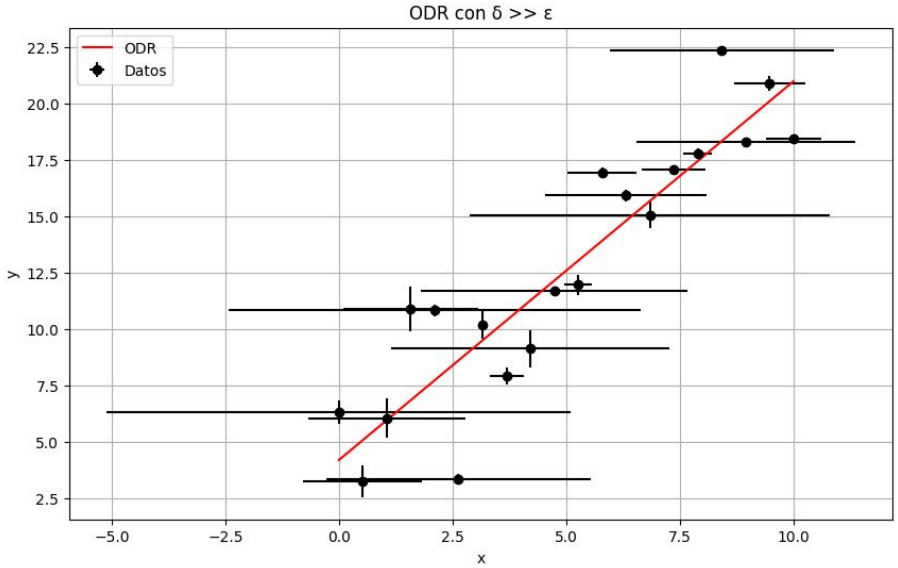
$$w_i = \frac{1}{\sigma_{\epsilon_i}} \; y \; d_i = \frac{\sigma_{\epsilon_i}}{\sigma_{\delta_i}} \qquad \text{con} \qquad w_i \geq 0 \;\; y \;\; d_i > 0$$

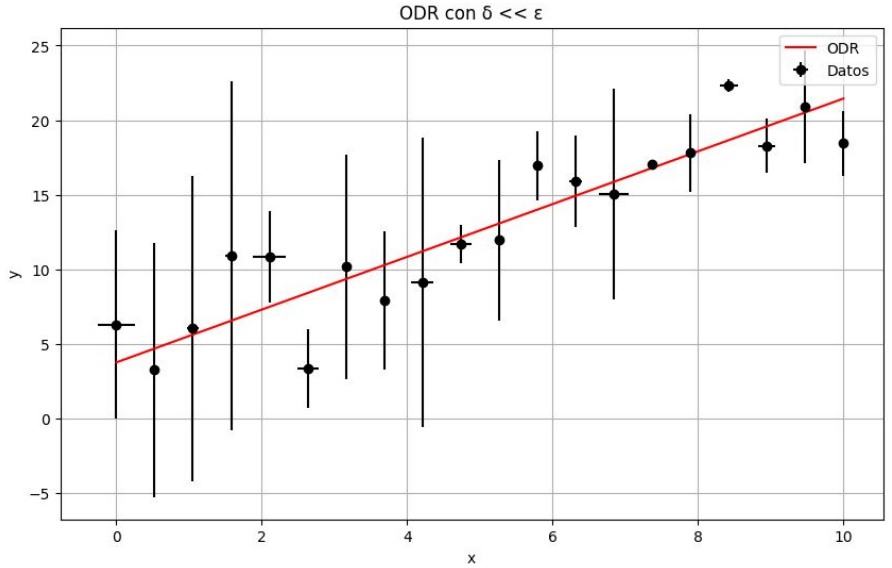
Si los errores no dependen sólo de la variación estándar esto “trae problemas”...

$$\Rightarrow \quad \min_{\beta_i, \delta_i} \sum_{i=1}^{n} w_i^2 \left\{ \left[f\left(X_i + \delta_i; \beta\right) - Y_i \right]^2 + d_i^2 \delta_i^2 \right\}$$

Resolviendo esta minimización se obtienen los parámetros β para el ajuste.

Introducción





La consideración de los pesos de los errores funciona de forma tal que desestima los errores cuanto mayores sean.

Simulaciones

Simulaciones: Noción sobre el algoritmo base

Muchos paquetes de ODR actuales (como el implementado en scipy) utilizan como base el ODRPACK creado por Boggs, Byrd y Schnabel.

Llamamos →

$$\begin{aligned} g_i(\tilde{\beta},\tilde{\delta}) &= w_i\left[f(X_i+\tilde{\delta}_i;\tilde{\beta})-Y_i\right] && i=1,\ldots,n \\ g_{i+n}(\tilde{\beta},\tilde{\delta}) &= w_i d_i \tilde{\delta}_i && i=1,\ldots,n, \end{aligned}$$

Definimos →

$$\left\|g(\tilde{\beta},\tilde{\delta})\right\|^2 \equiv \sum_{i=1}^{2n}\left[g_i(\tilde{\beta},\tilde{\delta})\right]^2.$$

La minimización nos queda →

$$\min_{\tilde{\beta},\tilde{\delta}}\left\|g(\tilde{\beta},\tilde{\delta})\right\|^2 \equiv \min_{\tilde{\beta},\tilde{\delta}}\sum_{i=1}^{2n}\left[g_i(\tilde{\beta},\tilde{\delta})\right]^2$$

Llamamos →

$$\tilde{\theta} = \begin{pmatrix} \tilde{\beta} \\ \tilde{\delta} \end{pmatrix}$$

Reemplazando en **g** y tomando en cuenta el Jacobiano →

$$g(\tilde{\theta}) = (g_1(\tilde{\theta}), g_2(\tilde{\theta}), \ldots, g(\tilde{\theta})_{2n})',$$

$$J_{ij} = \frac{\partial g_i}{\partial \tilde{\theta}_j}.$$

Taylor a primera derivada →

$$g(\theta^c) + J(\theta^c)s \equiv g^c + J^c s$$

$$s = \tilde{\theta} - \theta^c.$$

Simulaciones: Noción sobre el algoritmo base

Obtenemos **S** de resolver

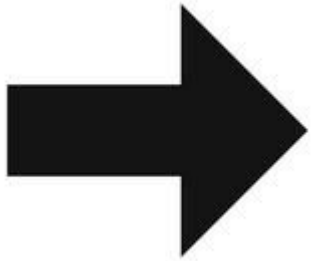
$$\min_s \|g^c + J^c s\|^2 \quad \text{subject to:} \quad \|s\| \leq \tau$$

Donde definimos **τ** como el intervalo de confianza para el cual el Taylor es una buena aproximación

Tras esto y al otorgar una aproximación inicial de los parámetros θ^c y de **τ** ;

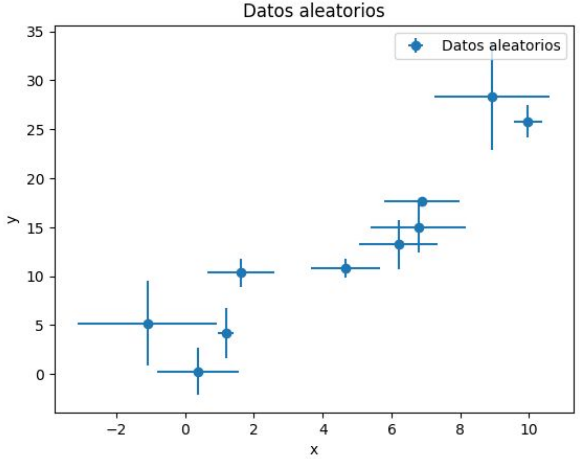
1) **Resolver la ecuación pasa S^c**
2) **$\theta^+ = S^c + \theta^c$**
3) **Si $||g(\theta^+)|| < ||g^c||$ eonces:**
   a) **$\theta^c = \theta^+$**
   b) **Ajusta τ si es necesario**
   c) **examina la convergencia**

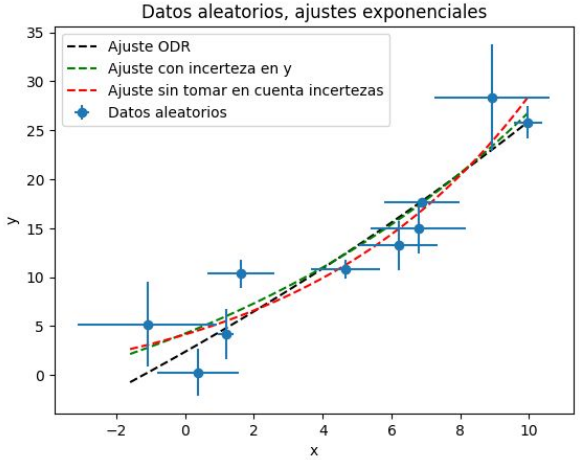
   sino:
   d) **reduce τ**
4) **Vuelve a empezar**



El algoritmo itera hasta que los cambios en los parámetros sean lo suficientemente chicos o hasta alcanzar el número máximo de iteraciones.

Simulaciones: función exponencial (ajuste lineal)





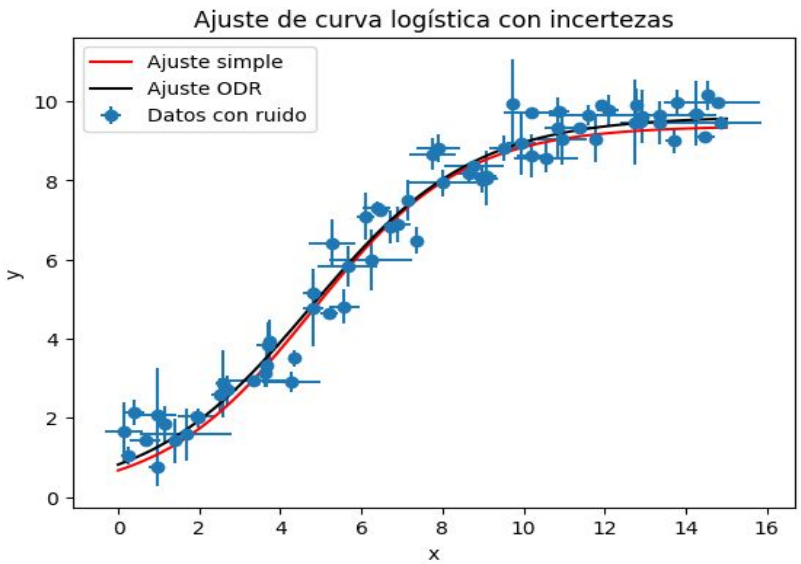
Ajuste exponencial un set de datos mediante tres métodos:

Podemos ver como actúan los pesos de las incertezas en los y_i en el caso de los cuadrados mínimos, y las incertezas en tanto en y_i como en x_i en el caso del método ODR.

Los ajustes se realizaron con la función $y=y_0+Ae^{x/t}$

Simulación no lineal (Curva logística)

$$f(x) = \frac{A}{1 + e^{-k(x - x_0)}}$$



	ODR	Cuadrados Mínimos	Real
A	9.617	9.377	10
ΔA	0.09	0.06	–
K	0.50	0.54	0.5
ΔK	0.03	0.02	–
x_0	4.76	4.77	5
Δx_0	0.14	0.11	–

En este ejemplo particular se observa lo siguiente:

- Los parámetros estimados por ODR son ligeramente más cercanos al valor real.
- Las incertezas son menores con los cuadrados mínimos. **Pero el motivo de esto es que estamos descartando los errores del eje x en el método de cuadrados mínimos, por lo que es razonable que esto suceda, independientemente de la efectividad del método.**

Simulación no lineal (Curva logística)

Para verificar los puntos mencionados se realizaron 1000 simulaciones de datos con tendencia de curva logística de los cuales se estimaron sus parámetro por medio de ambos métodos para compararlos.

De esto se obtuvo que:

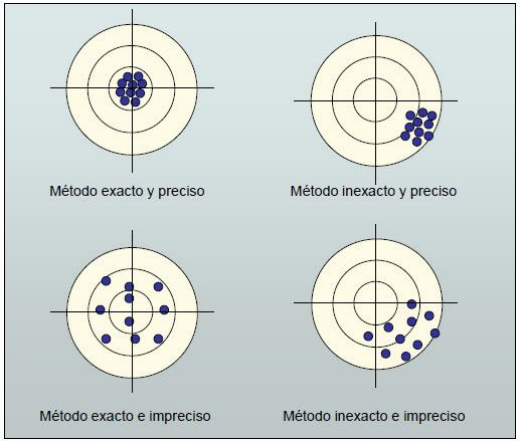
- El 75.2% de las simulaciones $|A^{ODR}-A|<|A^{\blacksquare MIN}-A|$.
- El 76.2% de las simulaciones $|K^{ODR}-K|<|K^{\blacksquare MIN}-K|$.
- El 75.7% de las simulaciones $|X_0{}^{ODR}-X|<|X_0{}^{\blacksquare MIN}-X|$.

→ Esto nos indica que los parámetros estimados por el método ODR son más exactos en una importante porción de los casos comparados a cuadrados mínimos.

- El 54.3% de las simulaciones $\Delta A^{\blacksquare MIN}<\Delta A^{ODR}$.
- El 66,7% de las simulaciones $\Delta K^{\blacksquare MIN}<\Delta K^{ODR}$.
- El 74.6% de las simulaciones $\Delta X_0{}^{\blacksquare MIN}<\Delta X_0{}^{\blacksquare MIN}$.

↓

Nuevamente las estimaciones de cuadrados mínimos parecieran ser más precisas **PERO, al comparar un método que utiliza errores en "x" e "y" vs uno que usa solo errores en "y" no es una comparación acertada.**



Comparación extendida para modelos no lineales

Comparación para modelos no lineales

- Se compara el sesgo, la varianza y el error cuadrático medio de los parámetros estimados, así como los estimados de la función usando ODR y OLS

- Se realiza con un modelo lineal, uno cuadrático, uno exponencial y uno sinusoidal

- Los errores δ y ϵ son generados como una distribución normal

- $w_i = 1 \quad ; \quad d_i = \dfrac{\sigma_{\epsilon_i}}{\sigma_{\delta_i}} = \{0.1\,;\,0.5\,;\,1.0\,;\,2.0\,;\,10.0\,;\,100.0\,;\,\infty\}$

Comparación para modelos no lineales

Si se conoce con exactitud la relación $d_i = \frac{\sigma_{\epsilon_i}}{\sigma_{\delta_i}}$

- Para sesgo de los parámetros:
  - ODR es mejor o igual que OLS en el 98% de los casos
  - ODR es claramente mejor en el 50% de los casos

- Para varianza y error cuadrático medio:
  - ODR es mejor o igual que OLS en el 98% de los casos
  - ODR es apreciablemente mejor en el 23% de los casos

- Para el modelo lineal no se encuentra una preferencia por ODR o OLS

Comparación para modelos no lineales

No se conoce con exactitud la relación $d_i = \frac{\sigma_{\epsilon_i}}{\sigma_{\delta_i}}$

➢ El error considerado es correcto dentro de un factor de 10

- Para sesgo de los parámetros:
  - ODR es mejor o igual que OLS en el 88% de los casos
  - ODR es significativamente mejor en el 59% de los casos

- Para varianza:
  - ODR es mejor o igual que OLS en el 90% de los casos
  - ODR es mejor en el 40% de los casos

- Para error cuadrático medio:
  - ODR es mejor o igual que OLS en el 87% de los casos
  - ODR es mejor en el 50% de los casos

Regiones/Intervalos de confianza

- Se obtienen de forma aproximada
- Se calcula la matriz de covarianza asintótica $\longrightarrow$ Intervalos aproximados
- Simulación de Monte Carlo en 4 problemas conocidos para valores de d_i conocidos y con error

Regiones/Intervalos de confianza

- Cuando se tiene d_i exacto, los estimados son buenos para las regiones e intervalos de confianza
- Los estimados de la región de confianza se degradan significativamente al considerar error en d_i
- Los intervalos de confianza son buenos cuando error considerado es correcto dentro de un factor de 2 y se degradan significativamente cuando es de un factor de 10

¿Preguntas?