Análisis de regresión

9 de septiembre
(semana 5)

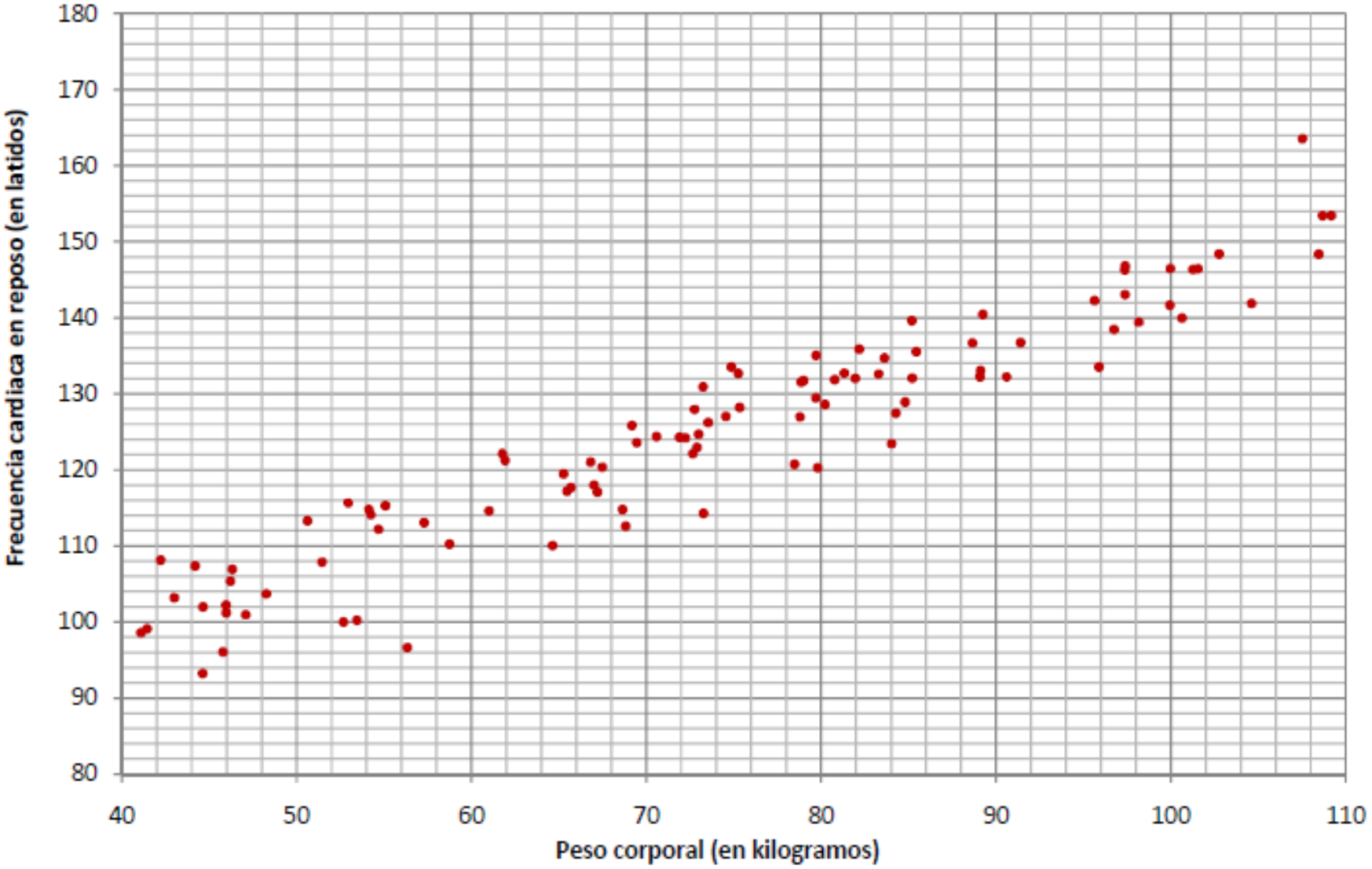
Plan de trabajo

1. Ejercicios aplicados de la regresión lineal simple
2. Otros temas de regresión lineal simple

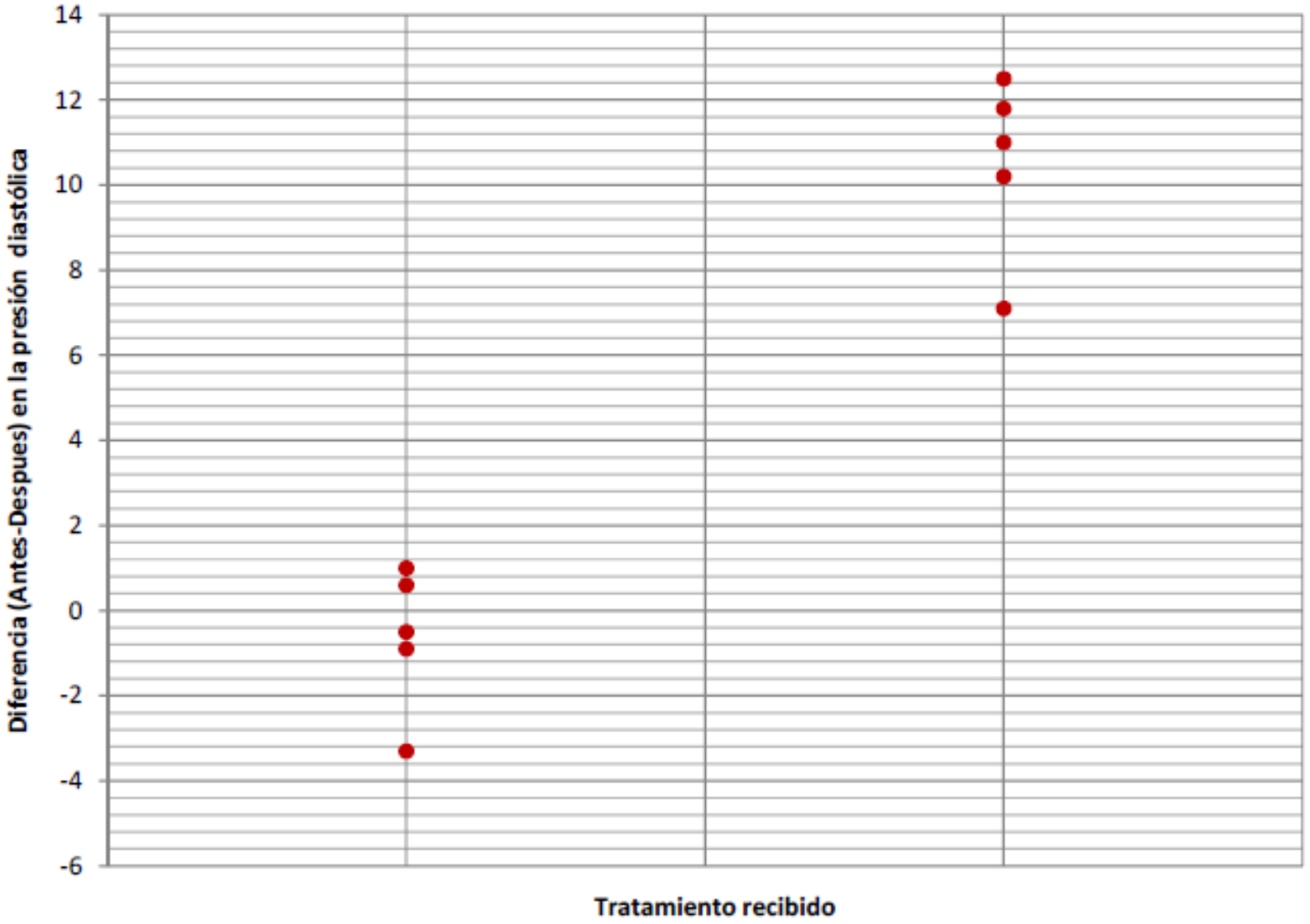
Nota:

- Esta semana, a más tardar el domingo, les entrego sus propuestas de proyecto (siguiente entrega: semana 11)
- Esta semana también subiré algunas soluciones a los problemas del taller 1.
- Por favor, impriman la hoja de teoremas que aparece en el drive.
- **Semana 6 (12 – 18 septiembre):**
  - Quiz por MOODLE. Vale el doble de un quiz de clase.
  - Sesión dudas pre-parcial 1 (viernes 16 o sábado 17)
- **Semana universitaria (19 – 25 septiembre):** ¿clase el miércoles?
- **Semana 7 (26 sept. – 2 octubre):** Parcial 1

Ejercicio en R (I)



Ejercicio en R (II)



Otros temas de regresión lineal simple

Inferencia exacta sobre la independencia de variables aleatorias normales

En el modelo de regresión simple **con errores normales**, es posible ver que

$$\hat{\beta}_1 = \hat{\rho}_{X,Y}\frac{S_Y}{S_x},\ \text{luego}\ \left\{T_C \middle| \mathbf{X}=\mathbf{x}\right\} = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}\left(\hat{\beta}_1\right)}} = \frac{\hat{\rho}_{X,Y}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2_{X,Y}}} \overset{H_0}{\sim} t\left(n-2\right),$$

en el sistema: $\begin{cases} H_0 : \beta_1 = 0 \therefore \rho_{X,Y} = 0 \text{ (independencia)} \\ \textit{versus} \\ H_1 :\ \beta_1 \neq 0 \therefore \rho_{X,Y} \neq 0 \text{ (no independencia)} \end{cases}$.

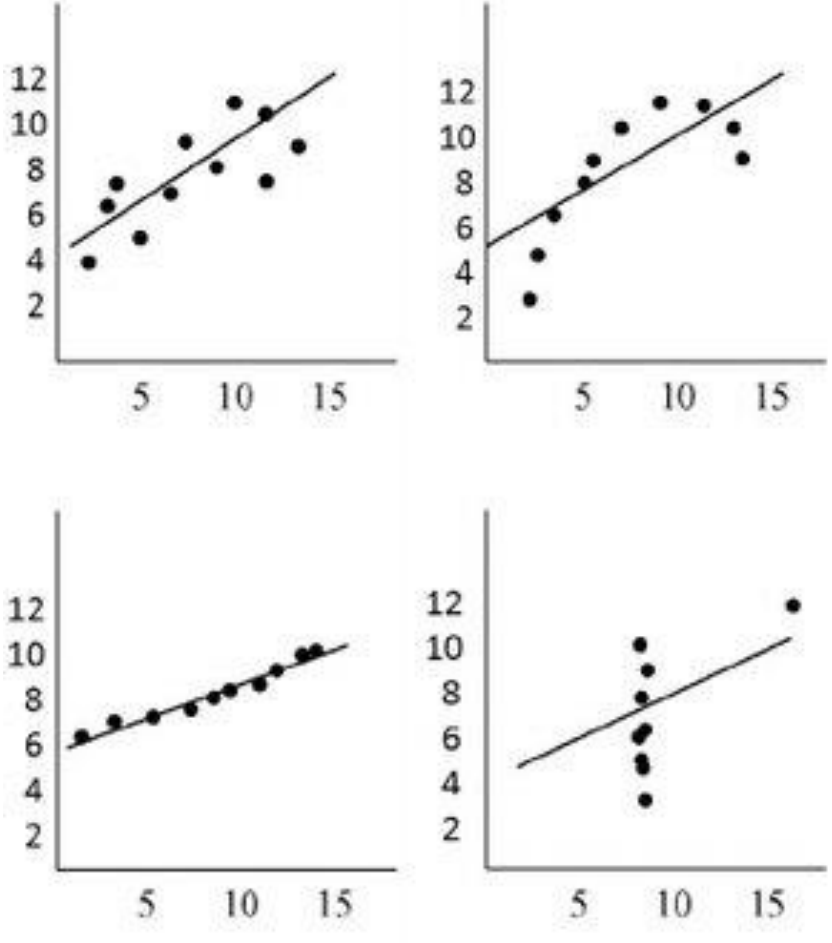
Como la distribución obtenida no depende de $\mathbf{X}=\mathbf{x}$, entonces, esta misma distribución es la distribución incondicional; así que el test de independencia puede ser juzgado como τ:"Rechazar H_0 si $\left|t_C\right| > t_{1-\frac{\alpha}{2}}\left(n-2\right)$"

- Esta prueba también sirve para vectores aleatorios normales bivariados.

Importancia de la exploración gráfica en la regresión lineal simple

Fuente: https://www.researchgate.net/publication/285672900_The_general_theory_of_culture_entrepreneurship_innovation_and_quality-of-life_Comparing_nurturing_versus_thwarting_enterprise_start-ups_in_BRIC_Denmark_Germany_and_the_United_States/figures?lo=1

Anscombe's Quartet



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Definición general de modelo lineal

Modelo lineal

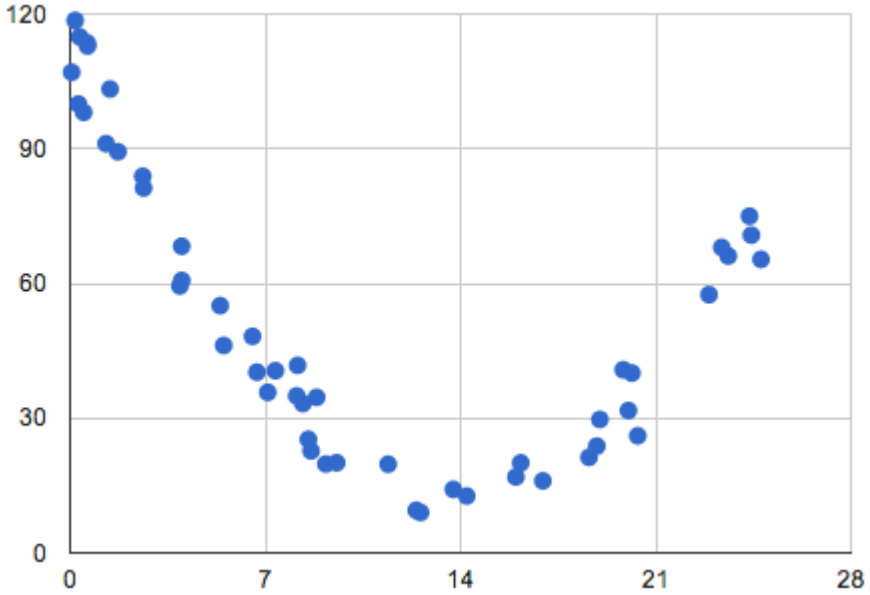
Considere una colección de v.a.s $Y_1, Y_2, \ldots, Y_n$ tales que:

$$\begin{cases} Y_i = \underset{\text{Compon. sistemática}}{\mu_i} + \underset{\text{Compon. aleatoria}}{\varepsilon_i}, \; i = 1, 2, \ldots, n. \\ \mu_i = E\left[Y_i\right] = g\left(\boldsymbol{\beta}, \mathbf{x}_i\right), \; i = 1, 2, \ldots, n, \\ \left\{\varepsilon_i\right\} \overset{\text{m.a.}}{\sim} N\left(0, \sigma^2\right) \end{cases}$$

El modelo se considera lineal siempre y cuando se cumplan las condiciones anteriores y además

$\dfrac{\partial \mu_i}{\partial \boldsymbol{\beta}}$ no depende de $\boldsymbol{\beta}$ en ninguna de sus componentes.

Ejemplo



Fuente:
https://www.statisticshowto.com
/quadratic-regression/

Modelo “linealizable”

Modelo linealizable

Considere una colección de v.a.s $Y_1, Y_2, \ldots, Y_n$ tales que:

$$\begin{cases} Y_i = h\left(\mu_i, \varepsilon_i\right), \ i = 1, 2, \ldots, n. \\ \mu_i = E\left[Y_i\right] = g\left(\boldsymbol{\beta}, \mathbf{x}_i\right), \ i = 1, 2, \ldots, n, \\ \left\{\varepsilon_i\right\} \overset{\text{m.a.}}{\sim} \end{cases}$$

El modelo se considera linealizable siempre y cuando se pueda llevar a la forma de un modelo lineal mediante una transformación de la variable respuesta o una reparametrización.

Modelo “linealizable” (ejemplos)

¿Cuáles de estos modelos son linealizables?

a) $Y_i = \beta_0 x_i^{\beta_1} \varepsilon_i$

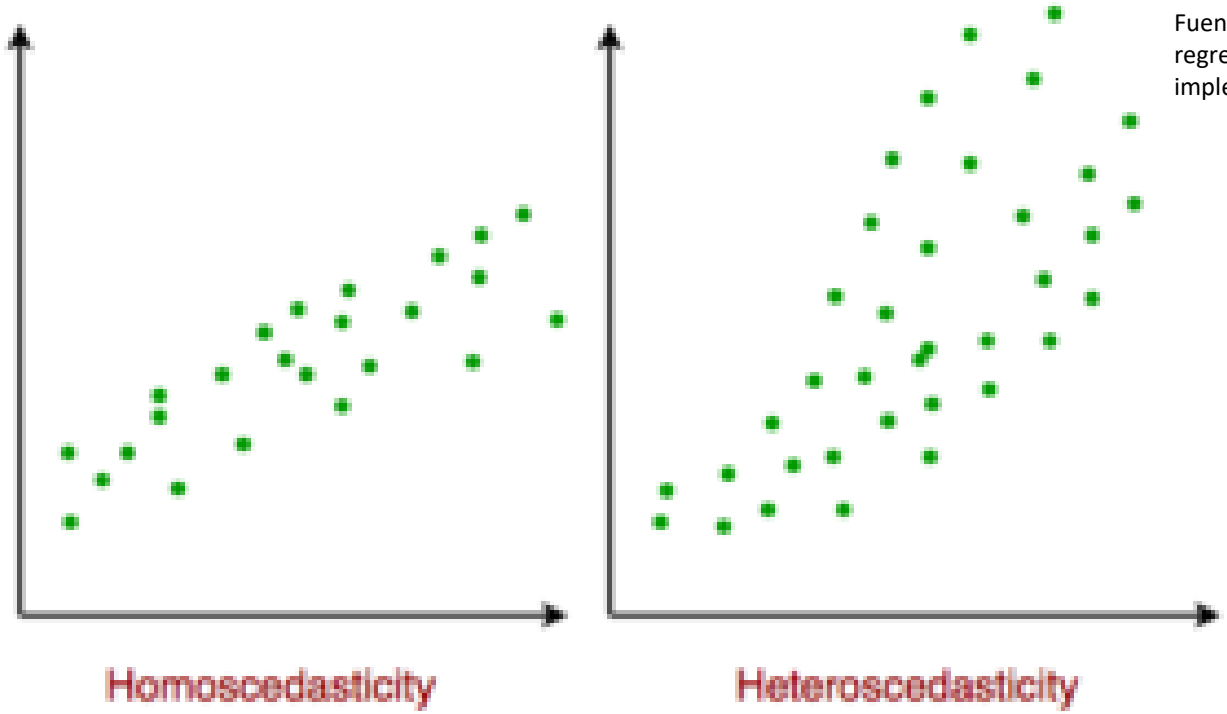
b) $Y_i = \left[1 + \exp\left(\beta_0 + \beta_1 x_i + \varepsilon_i\right)\right]^{-1}$

c) $\ln Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$

d) $Y_i = \dfrac{\beta_0 + x_i^2}{\beta_1} + \varepsilon_i$

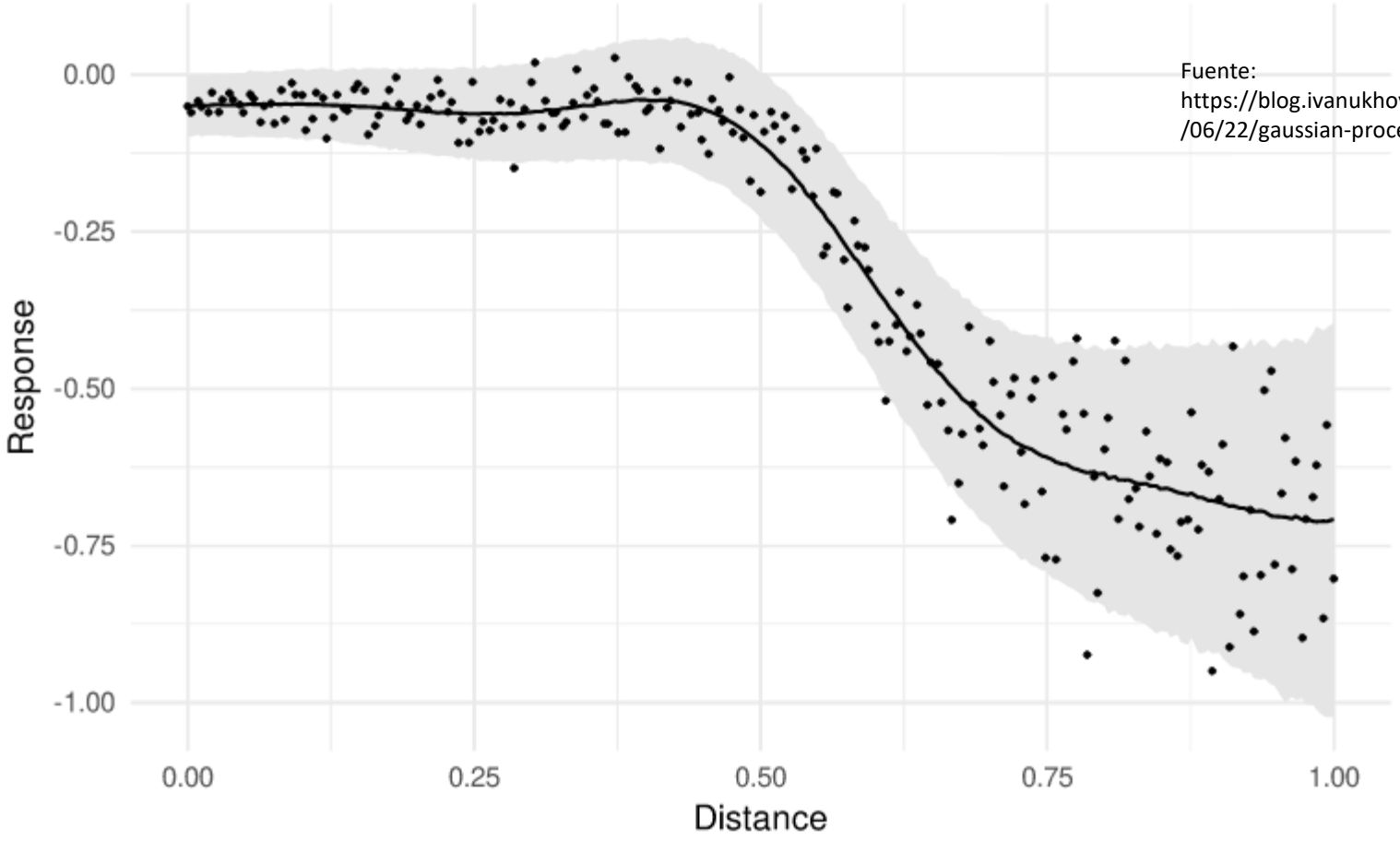
- Hay que ser muy cuidadoso con la evaluación de supuestos de esos modelos, con la manera en que se predice la variable dependiente original y con las propiedades de los estimadores.

Problema de heteroscedasticidad



Fuente: https://prutor.ai/linear-regression-python-implementation/

Problema de heteroscedasticidad



Fuente:
https://blog.ivanukhov.com/2020/06/22/gaussian-process.html

Estimación de los coeficientes de regresión

❖ **Parámetros:** β_0 : intercepto, β_1 : pendiente

Método de mínimos cuadrados ponderados: Suponga que se tiene el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 g(x_i) + \varepsilon_i; \quad 1 \le i \le n$$

Donde:

$$1.\ E(\varepsilon_i) = 0 \ \forall i$$

$$2.\ Var(\varepsilon_i) = \sigma^2 f(x_i) \ \forall i$$

$$3.\ Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$$

El estimador de mínimos cuadrados ponderados es:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{\beta_0, \beta_1} Q_\omega(\beta_0, \beta_1) = \sum_{i=1}^{n} \omega_i \left(y_i - \beta_0 - \beta_1 g(x_i)\right)^2, \ \omega_i \propto \frac{1}{Var(\varepsilon_i)}$$

Regresión resistente a valores extremos

- El modelo de regresión lineal simple funciona muy bien y permite hacer una gran cantidad de procedimientos inferenciales, siempre y cuando se cumplan los supuestos. Sin embargo, este procedimiento es bastante simple a valores (extremos) atípicos en una de las variables.

- Por ello, algunos autores han estudiado el uso de otras funciones o criterios para optimizar la búsqueda de las estimaciones:

$$\arg\min_{\beta_0,\beta_1} L(\beta_0,\beta_1) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \quad \textbf{LAD (Least absolute deviation) regression}$$

$$\arg\min_{\beta_0,\beta_1} L(\beta_0,\beta_1) = \max_{1\le i\le n} |y_i - \beta_0 - \beta_1 x_i| \quad \textbf{Minimax deviation regression}$$

$$\arg\min_{\beta_0,\beta_1} L(\beta_0,\beta_1) = \sum_{i=1}^{n} d^2\left((x_i,y_i),\text{recta}\right) \quad \textbf{Orthogonal (Deming) regression}$$

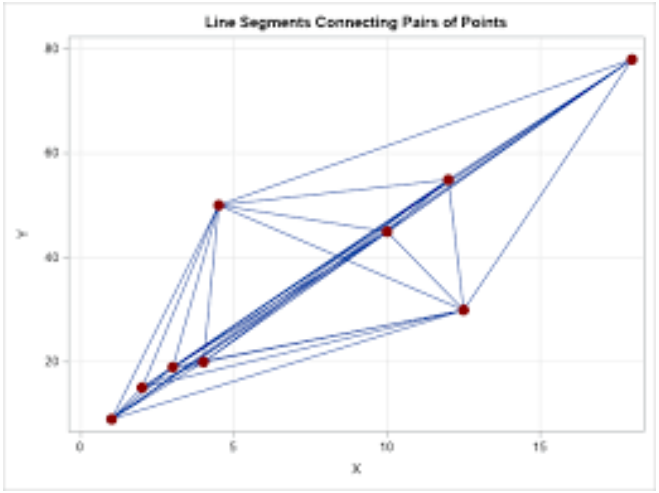
Regresión resistente a valores extremos (II)

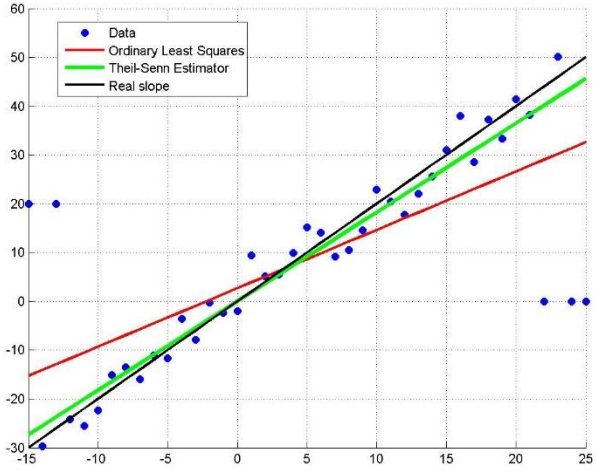
- Otra alternativa sería buscar modelar un aspecto resistente a valores extremos, como las medianas. De allí nace la regresión de Theil-Sen.

- Este método consiste en utilizar como estimación de la pendiente a la mediana de todas las pendientes:

$$\hat{\beta}_1 = \text{med}\left\{\frac{y_j - y_i}{x_j - x_i} : i \neq j\right\}$$

$$\hat{\beta}_0 = \text{med}\left\{y_j - \hat{\beta}_1 x_j\right\}$$





Fuente: https://ww2.mathworks.cn/matlabcentral/fileexchange/34308-theil-sen-estimator?s_tid=FX_rc2_behav

Regresión resistente a valores extremos (III)

- Este método produce estimadores insesgados de la pendiente y del intercepto. Además, acepta un porcentaje de corrupción de alrededor de 29.3% de datos atípicos. Exploren la librería `mblm` de `R`.

- Este y los otros métodos tienen la desventaja de que no tienen tanta teoría desarrollada como el modelo de regresión lineal clásico. Además, se han desarrollado mecanismos para identificar posibles inconvenientes que puedan afectar negativamente al modelo de regresión lineal clásico.