Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

Datos como realizaciones de una variable aleatoria

- Un **conjunto de datos** puede verse como *realizaciones independientes* de una variable aleatoria.
- Ejemplo: $X \sim \mathcal{N}(0, 1)$.
- Cada nueva observación agrega información, pero también refleja la **variabilidad inherente**.

Cuando un dato no encaja...

Smita (2020)

"In many contexts, most of the data come from the same generating distribution. However, occasionally observations appear that respond to a *different nature*."

- El **ruido** es parte natural de la variabilidad de la variable aleatoria.
- Los **outliers genuinos** pueden provenir de otro mecanismo generador.
- La tarea estadística: **distinguir** entre
  1. observaciones que parecen inusuales pero son parte del mismo proceso, y
  2. observaciones que de hecho provienen de otro mecanismo.

Ruido *vs.* outliers

Ruido: variabilidad aleatoria inherente al fenómeno y *parte del modelo probabilístico*.

En regresión lineal:

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

El término ε *representa el ruido*. No es "basura"; cuantifica incertidumbre no explicada.

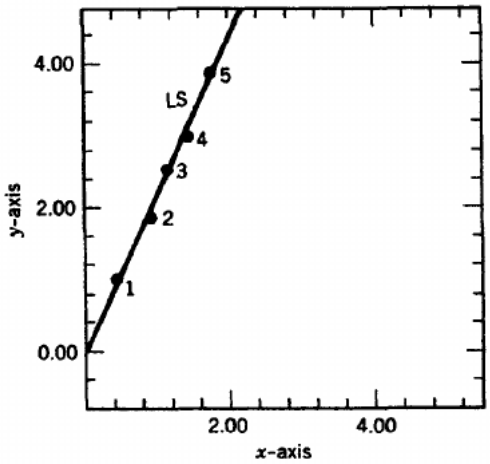
Outliers: observaciones que se desvían de tal forma que sugieren un *mecanismo generador distinto*.

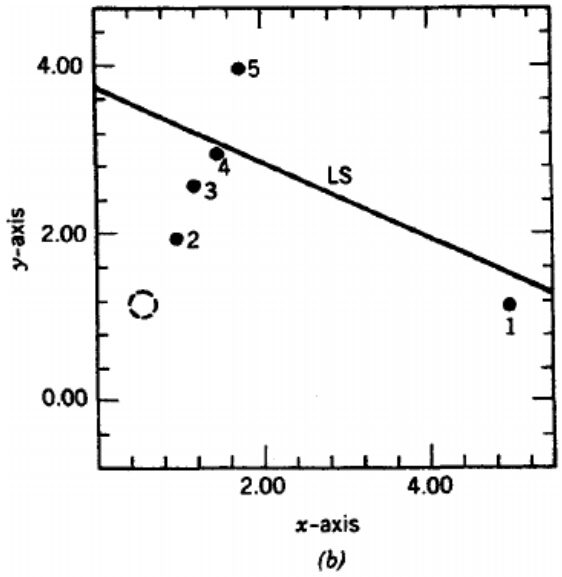
Definición clásica (Hawkins, 1980)

"Any applied statistician who has analysed a number of sets of real data is likely to have come across 'outliers'. The intuitive definition of an outlier would be 'an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism'."

Tipos de outliers (visión operativa)

- **Globales**: puntos aislados lejos del patrón general.
- **Contextuales**: atípicos dado un contexto (p. ej., 30°C en invierno).
- **Colectivos**: grupos de observaciones anómalas en conjunto.


4.00
2.00
0.00
y-axis
LS
1
2
3
4
5
2.00
4.00
x-axis



(b)

Modelo lineal clásico

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad \hat{\beta} = (X^\top X)^{-1} X^\top y.$$

- El **ruido** está modelado explícitamente vía ε.
- **Candidatos a outliers** aparecen cuando observaciones se apartan más de lo esperable bajo el supuesto $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Predicciones en regresión lineal

En mínimos cuadrados ordinarios:

$$\hat{y} = X\hat{\beta}, \quad \hat{\beta} = (X^\top X)^{-1} X^\top y.$$

Puede escribirse como:

$$\hat{y} = Hy,$$

donde

$$H := X(X^\top X)^{-1} X^\top$$

es la **Hat Matrix**.
Se asume que $X^t X$ es invertible.
La **Hat Matrix** transforma los las observaciones y en su estimador por mínimos cuadrados.

Propiedades de la Hat Matrix

- H es simétrica e idempotente: $HH = H$.
- Los elementos diagonales h_{ii} se llaman **leverages**.
- Interpretación: h_{ii} mide cuánta *auto-influencia* tiene la observación i sobre su valor ajustado $\hat{y}_i$.
- Suma de los leverages: $\sum_i h_{ii} = p$, número de parámetros (incluido el intercepto).

Leverage como herramienta de diagnóstico

- Observaciones con h_{ii} alto son **potencialmente influyentes**.
- No siempre tienen residuos grandes, pero pueden "arrastrar" la recta.
- Reglas prácticas comunes:
  - $h_{ii} > 2p/n$ (posible punto de interés).
  - $h_{ii} > 3p/n$ (considerar como alto leverage).
- El leverage permite identificar **outliers en el espacio de predictores** (no necesariamente en y).

Residuos y leverage

- Un punto puede tener:
  - Residuo grande $\Rightarrow$ posible **outlier vertical**.
  - Leverage grande $\Rightarrow$ punto con mucha influencia potencial.
- El análisis conjunto de **residuos estandarizados** y **leverages** permite distinguir:
  1. Outliers "genuinos" (residuo grande).
  2. Puntos influyentes (alto leverage).

Idea general

- Los **Single-Case Diagnostics** analizan el impacto de eliminar una sola observación del ajuste.
- Permiten cuantificar:
  1. **Influencia** de cada observación en los parámetros.
  2. **Sensibilidad** de las predicciones y residuos al remover un dato.
- Herramientas clave:
  - Distancia de Cook.
  - DFFITS.
  - DFBETAS.

Distancia de Cook

- Resume el efecto de eliminar la observación i en todos los coeficientes.
- Definición:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p\,\hat{\sigma}^2},$$

  donde $\hat{y}_{j(i)}$ son las predicciones al eliminar el caso i.
- Interpretación:
  - D_i grande $\Rightarrow$ observación influyente en el modelo.
  - Reglas prácticas: $D_i > 1$ (en modelos pequeños), o $D_i > \frac{4}{n}$ (como criterio general).

DFFITS

- Mide el efecto de eliminar el dato i en su *propia predicción*.
- Fórmula:

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}},$$

  donde $\hat{y}_{i(i)}$ es la predicción de y_i al omitir el caso i.
- Regla de decisión:

$$|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}} \quad \Rightarrow \quad \textit{posible caso influyente}.$$

DFBETAS

- Evalúan el cambio relativo en cada coeficiente β_j al remover una observación:

$$\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\hat{\sigma}_{(i)}\sqrt{c_{jj}}},$$

  donde c_{jj} es el j-ésimo elemento diagonal de $(X^\top X)^{-1}$.
- Interpretación:
  - Valor grande $\Rightarrow$ observación influye de manera importante en el coeficiente β_j.
  - Regla práctica: $|\text{DFBETAS}_{ij}| > \frac{2}{\sqrt{n}}$.

Resumen visual

- **Cook's distance**: efecto global en todos los coeficientes.
- **DFFITS**: impacto en la predicción del mismo dato.
- **DFBETAS**: impacto en cada parámetro individual.

Referencias breves

[1] D. M. Hawkins (1980).
Identification of Outliers.
Chapman & Hall.

[2] P. J. Rousseeuw, A. M. Leroy (1987).
Robust Regression and Outlier Detection.
Wiley.

[3] A. Smiti (2020).
A critical overview of outlier detection methods.
Computer Science Review, vol. 38, article 100306.