# Erregresio-analisia

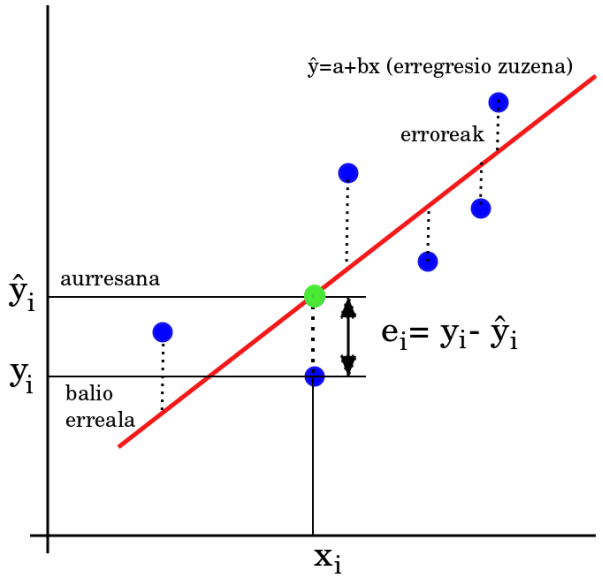
Josemari Sarasola

Estatistika eta datuen analisia

Gizapedia

## Helburua

$(x_i, y_i)$ datuei, bi aldagaiko puntu-hodei bati alegia, zuzen bat doitzea, puntuetara gehien hurbiltzen dena. $y$ aldagaitzat aldagai dependentea hartzen da.



## Nola eratu zuzena?

Errore karratu guztien batura txikien egiten duen zuzena eratuko da, karratu txikienen zuzena deitzen dena. Beraz, zuzena $\hat{y} = a + bx$ izanik, problema hau da:

$$\min_{a,b} \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

## Nola eratu zuzena?

Horretarako, minimotu beharrekoa $a$ eta $b$ parametroei buruz deribatu eta 0-ra berdindu behar da. Horrela, erregresioaren *ekuazio normalak* eskuratzen dira:

$$\frac{\partial \sum_i (y_i - a - b x_i)^2}{\partial a} = 0 \rightarrow \sum_i y_i = na + b \sum_i x_i$$
$$\frac{\partial \sum_i (y_i - a - b x_i)^2}{\partial b} = 0 \rightarrow \sum_i x_i y_i = a \sum_i x_i + b \sum_i x_i^2$$

## Nola eratu zuzena?

Aurreko bi ekuazioak ebatziz, honela zenbatetsi edo estimatzen dira $a$ eta $b$ parametroak:

$$b = \frac{s_{xy}}{s_x^2}$$

$$a = \overline{y} - b\overline{x}$$

10. ariketa

## Doikuntzaren egokitasuna

Karratu txikienen erregresioan berdintza hau betetzen da:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

$s_y^2$ bariantza totala da, erregresio-zuzenak $x$ aldagairen bitartez esplikatzea helburu duguna. $s_{\hat{y}}^2$ bariantza azaldua da, $x$-rekiko erregresioaren bitartez azaltzen dugun $y$ aldagaiaren bariantzaren zatia, eta $s_e^2$ erroreen bariantza, $x$-rekiko erregresioaz azaltzen ez dugun zatia. Beraz:

*bariantza totala=bariantza azaldua+azaldu gabeko bariantza*

## Doikuntzaren egokitasuna

Beraz, bariantza totala bati buruz, bariantza azaldua zenbat eta handiagoa izan, erregresio-zuzena orduan eta hobeto doituko da datuetara. Hargatik, doikuntzaren egokitasun neurri gisa *mugatze-koefizientea* izenekoa eratzen da:

$$R^2 = \frac{s^2_{\hat{y}}}{s^2_y} = 1 - \frac{s^2_e}{s^2_y}$$

## Doikuntzaren egokitasuna

$[0, 1]$ tarteko balioak hartzen ditu eta honelako erregela erabil daiteke interpretaziorako, lagin-errorearen eta antzeko ikerketen erreserbapean:

- $0 < R^2 < 0.3 \rightarrow$ doikuntza eskasa;
- $0.3 < R^2 < 0.6 \rightarrow$ doikuntzaren kalitatea ertaina;
- $R^2 > 0.6 \rightarrow$ doikuntzaren egokitasun handia;
- $R^2 = 1 \rightarrow$, doikuntza erabatekoa: puntu guztiak lerrokatuta daude.

## Doikuntzaren egokitasuna

Erregresio-zuzenaren bi propietate hauek erraztu egiten dute mugatze-koefizientearen kalkulua:

- $\sum_i e_i = 0 \rightarrow \overline{e} = 0 \rightarrow s_e^2 = \frac{\sum_i e_i^2}{n}$;
- $\overline{y} = \overline{\hat{y}}$.

11. ariketa: 1, 2 atalak

## Erregresioaren diagnosia: errore-diagramak

Karratu txikienen erregresioa garatu dugun bezala garatzeko, zenbait baldintza bete behar da. Baldintza horiek betetzen diren, erregresio-ereduaren diagnosia egiteko alegia, errore-diagrama eratu eta azter daiteke.

## Erregresioaren diagnosia: errore-diagramak

**Errore-diagrama** kartesiar-diagrama besterik ez da, abzisa-ardatzean $x_i$ aldagaiaren balioak eta ordenatu-ardatzean $e_i$ erroreenak jartzen dituena.

## Nola interpretatu errore-diagrama?

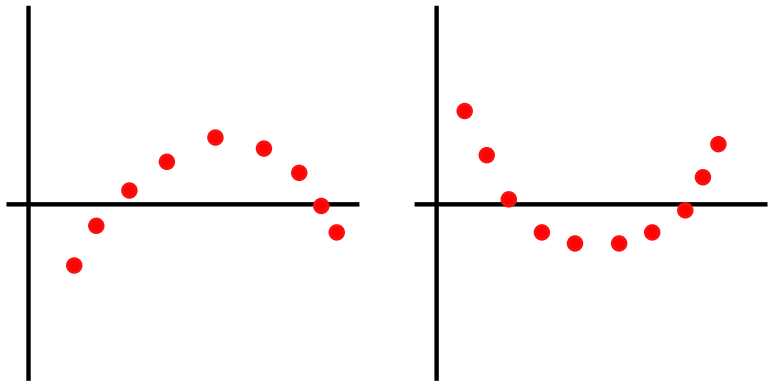


Figure: **Zehaztapen-errorea**: lerro desegokia aukeratu da; adibidez, zuzenaren ordez, kurba bat doitzea litzateke egokiena. Zantzu txarra da, beraz.

## Nola interpretatu errore-diagrama?

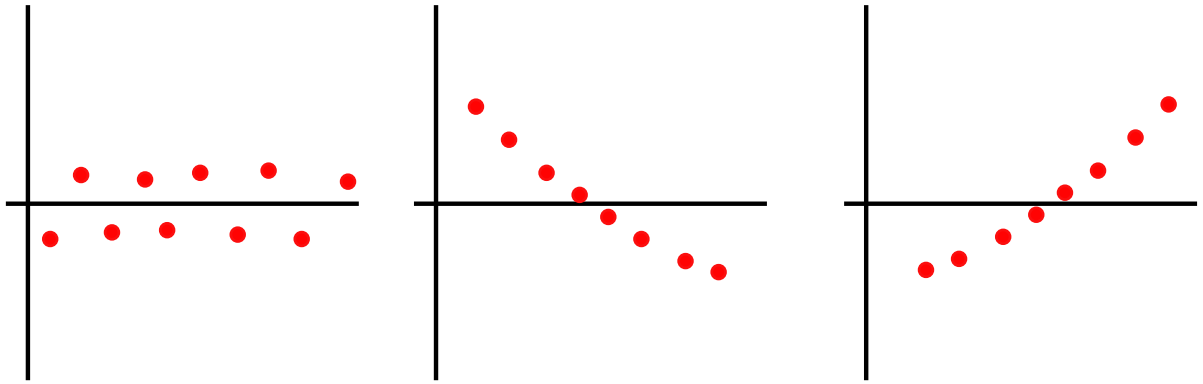


Figure: **Autokorrelazioa**: ondoz ondoko erroreak loturik daude. Zantzu txarra da.

## Nola interpretatu errore-diagrama?

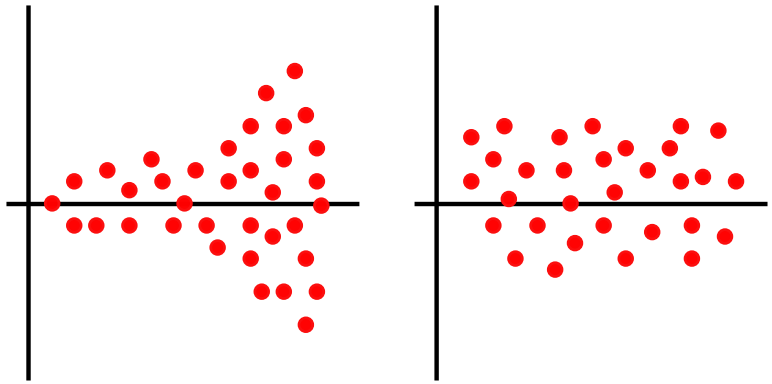


Figure:

- Erroreen sakabanatze desberdina: **heteroskedastikotasuna** (zantzu txarra).
- Erroreen sakabanatze berdina: **homoskedastikotasuna** (zantzu ona).

## Nola interpretatu errore-diagrama?

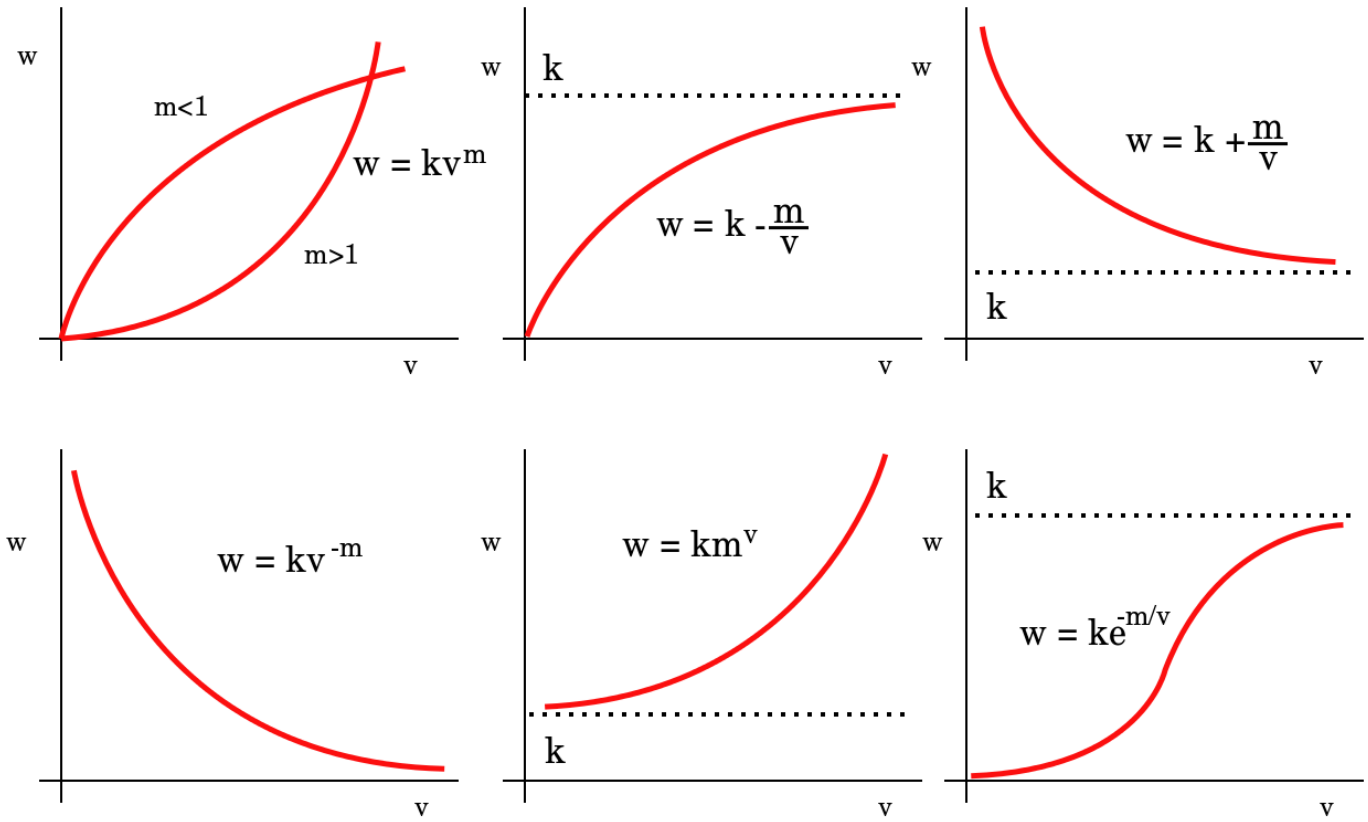
Erregresio-eredua ontzat emateko, errore-diagraman erroreak zoriz edo modu kaotikoan banatuak izan behar dira orokorrean, joera garbirik gabe.

11. ariketa: 3. atala

## Eredu ez-linealak

Erregresio-zuzen bat doitzen ikasi dugu. Nola doitu erregresio-kurba bat?
Orokorrean, kurba linealizatu egingo dugu, zuzen bihurtu alegia, eta zuzeneko $a$ eta $b$ zenbatetsi ondoren, kurbara itzuliko gara ostera.

## Eredu ez-linealak: kurba-katalogoa



## Eredu ez-linealak: katalogoko kurben linealizazioa

| Lerroa | Linealizazioa ($y = a + bx$) | a | b |
|---|---|---|---|
| $w = kv^m$ | $\ln w = \ln k + m \ln v$ | $a = \ln k$ | $b = m$ |
| $w = k - \frac{m}{v}$ | $w = k + (-m)\frac{1}{v}$ | $a = k$ | $b = -m$ |
| $w = k + \frac{m}{v}$ | $w = k + m\frac{1}{v}$ | $a = k$ | $b = m$ |
| $w = kv^{-m}$ | $\ln w = \ln k + (-m) \ln v$ | $a = \ln k$ | $b = -m$ |
| $w = km^v$ | $\ln w = \ln k + (\ln m)v$ | $a = \ln k$ | $b = \ln m$ |
| $w = ke^{-\frac{m}{v}}$ | $\ln w = \ln k + (-m)\frac{1}{v}$ | $a = \ln k$ | $b = -m$ |

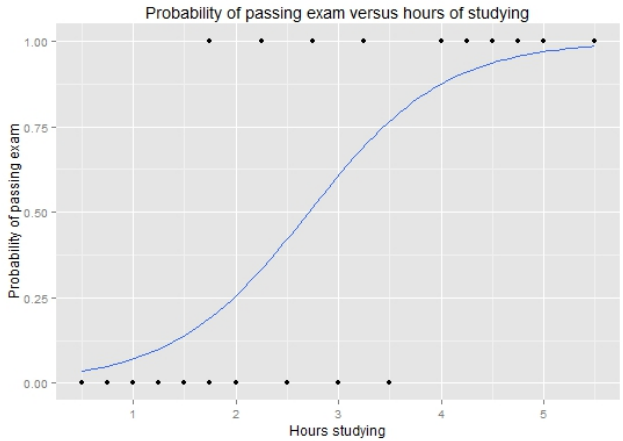
## Eredu ez-linealak: mugatze-koefizientearen kalkulua

Erregresio-kurba baten mugatze-koefizientea forma linealizatuarekin kalkulatu behar da, alegia, $y$ balio errealak, aurresanak eta erroreak zuzenari buruz kalkulatu behar dira. Hala egiten ez bada, mugatze-koefizientea ez da [0,1] tartera mugatuko.

12. eta 13. ariketak

## Logit eredua

**Logit eredua** edo **erregresio logistikoa** erregresio-kurba berezia da, non aldagai independentea *dosi* izeneko aldagai bat den, eta dependentea, zerbait gertatzeko *probabilitatea* adierazten duena. Beraz, aldagai dependentea probabilitatea denez, bere balioak [0-1] tartera mugatuak behar dira izan.
Ohiko eran, zenbat eta dosi handiagoa, orduan eta probabilitate handiagoa dago.



## Logit eredua

Aurreko adibidean kurba logistiko tipiko bat azaltzen da. Ordenatu-ardatzean balioak [0,1] tartera mugatzen dira, probabilitate batez ari garenez gero. Bestalde, probabilitate-gehikuntza gehienak dosi-kopuru tarte batean gertatzen dira. Dosi oso txikietarako edo handietarako probabilitatearen aldaketa oso txikia da (adibidez, muga batetik aurrera, asko estudiatuta ere ez duzu aprobatzeko probabilitatea askoz ere handiagoa izango).

## Logit eredua

$x$ dosia eta $p$ probabilitate izanik, hau da kurba logistikoaren forma linealizatua:

$$\ln\left(\frac{p}{1-p}\right) = a + bx$$

Bestalde, edozein erregresio-kurbatan bezalaxe, mugatze-koefizientea aurreko forma linealizatuan oinarrituta kalkulatu behar da.