5.Gaia: Erregresioa

Cristina Alcalde - Arantxa Zatarain

Donostiako Unibertsitate Eskola Politeknikoa - UPV/EHU

Bi dimentsiotako aldagaiak

Multzo bateko osagaien bi atributu aztertzen ditugunean bi dimentsiotako aldagaia dugu. n elementuz, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, osatutako laginaren osagaiak elkarzut diren ardatz sisteman adierazten ditugunen sakabanatze-diagrama lortzen dugu, bi atributuen arteko dependentzia estatistikoa ager daiteke eta hortik erlazio funtzionala atera daiteke.

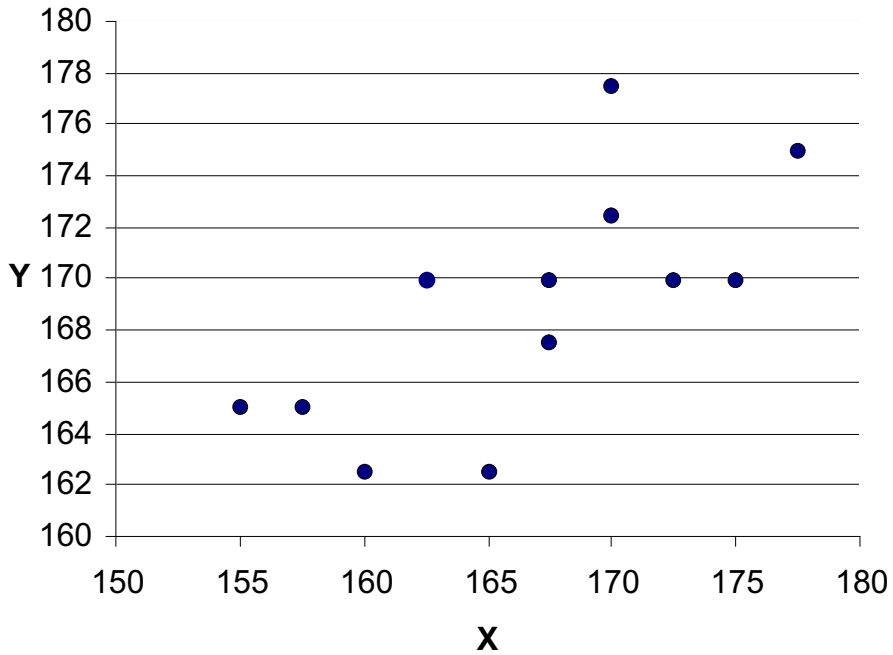
Adibidea

Aita-semea edo ama-alaba eran osatutako 12 bikoteen altuerak ditugu taulan.

X	Y
162.5	170
157.5	165
167.5	170
160	162.5
170	172.5
155	165
175	170
165	162.5
170	177.5
167.5	167.5
172.5	170
177.5	175

Adibidea

Datu horien adierazpen grafikoa egiten dugunean dispertsio diagrama lortzen dugu.



Doiketa lortzeko prozesua

Demagun datu kopurua infinitua dela eta x balio bakoitzari dagozkien y balioen batezbestekoa $E(x, y)$ dela, $(x, E(x, y))$ puntu guztietatik pasatzen den lerroa x-gaineko y-ren erregresio lerroa da.
Lerro hori ezagutzen denean x balioentzat y balioa lor dezakegu.
Erregresioa aztertzerakoan hiru pausu egin behar ditugu:

1. Erregresio-lerroaren mota aukeratu. Hau puntu-multzoa edo sakabanatze-diagrama aztertuz egiten da.
2. Erregresio-lerroaren parametroak lortu. Parametroak lortzeko karratu minimoen metodoa erabiliko dugu.
3. Doiketaren kritika. Parametroak beste lagin batez lortzen direnean aldaketa handiak badira lortutako erregresio-lerroa ez da praktikoa izango.

Karratu minimoen metodoa

Demagun $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ laginaren puntu-multzoa dela eta $y = f(x)$ funtzioa aukeratu dugula erregresio-lerroaren adierazpen analitikoa bezala, funtzio horretan parametro batzuk azalduko dira, adibidez, $y = a + bx$ zuzena aukeratzen badugu bi parametro finkatu behar ditugu, $y = ax^2 + bx + d$ parabolaren kasuan 3 parametro, eta bi parametro lortu beharko genituzke $y = ae^{bx}$ esponentzialaren kasuan eta $y = ax^b$ potentzialaren kasuan.

Karratu minimoen metodoa

Parametroak lortzeko

$$F(a, b, \ldots) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

funtzioa definitzen dugu, hau da lagineko y_i eta $f(x_i)$ funtzioaren bidez lortzen diren balioen arteko diferentzien karratuen batura.
Parametroen balioak lortzen ditugu $F(a, b, \ldots)$ funtzio hau minimoa izan dadin. Horretarako

$$\frac{\partial F}{\partial a} = 0, \frac{\partial F}{\partial b} = 0, \ldots$$

baldintzak bete behar dira. Lortutako ekuazio-sistema ebatziz parametroen balioak lortzen dira.

Erregresio lineala

Demagun erregresio-lerro bezala $y = a + bx$ zuzena aukeratu dugula. $F(a, b)$ funtzioa era honetan gelditzen zaigu,

$$F(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Funtzioa minimoa izan dadin

$$\frac{\partial F}{\partial a} = 0, \frac{\partial F}{\partial b} = 0$$

bete behar dira. Beraz,

$$\frac{\partial F}{\partial a} = 2 \sum_{i=1}^{n} [(y_i - a - bx_i)(-1)] = 0$$

$$\frac{\partial F}{\partial b} = 2 \sum_{i=1}^{n} [(y_i - a - bx_i)(-x_i)] = 0$$

Erregresio lineala

Bi ekuazio horietan banatze, elkartze eta trukatze propietateak erabiliz, sistema hau lortzen dugu

$$\begin{cases} \sum_{i=1}^{n} y_i & -\sum_{i=1}^{n} a & -\sum_{i=1}^{n} bx_i & = 0 \\ \sum_{i=1}^{n} (y_i \cdot x_i) & -\sum_{i=1}^{n} ax_i & -\sum_{i=1}^{n} bx_i^2 & = 0 \end{cases} \Longrightarrow$$

$$\begin{cases} na & +b\sum_{i=1}^{n} x_i & = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i & +b\sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} (y_i \cdot x_i) \end{cases} .$$

Sistema honen ekuazioak Gauss-en ekuazio normalak deitzen dira.

$$B_x = \sum_{1=1}^{n} x_i, \quad B_{xx} = \sum_{1=1}^{n} x_i^2$$

$$B_y = \sum_{1=1}^{n} y_i, \quad B_{xy} = \sum_{1=1}^{n} x_i \cdot y_i$$

Erregresio lineala

eta sistemaren soluzioa

$$a = \frac{\begin{vmatrix} B_y & B_x \\ B_{xy} & B_{xx} \end{vmatrix}}{\begin{vmatrix} n & B_x \\ B_x & B_{xx} \end{vmatrix}}$$

$$b = \frac{\begin{vmatrix} n & B_y \\ B_x & B_{xy} \end{vmatrix}}{\begin{vmatrix} n & B_x \\ B_x & B_{xx} \end{vmatrix}}$$

Kobariantza

$$b = \frac{\begin{vmatrix} n & B_y \\ B_x & B_{xy} \end{vmatrix}}{\begin{vmatrix} n & B_x \\ B_x & B_{xx} \end{vmatrix}} = \frac{nB_{xy} - B_x B_y}{nB_{xx} - (B_x)^2}$$

Frakzioaren izendatzailea eta zenbatzailea n^2-z zatituz

$$b = \frac{\frac{B_{xy}}{n} - \frac{B_x}{n}\frac{B_y}{n}}{\frac{B_{xx}}{n} - \frac{(B_x)^2}{n^2}}$$

Zenbatzailean dugun adierazpena bi dimentsioetako laginaren kobariantza deitzen da eta izendatzailean duguna x laginaren bariantza da.

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

Erregresio zuzena

b balioa y-ren x-kiko erregresio-koefizientea da.

$$a = \frac{B_y B_{xx} - B_x B_{xy}}{nB_{xx} - (B_x)^2}$$

b-ren balioa sistemaren lehenengo ekuazioan ordezkatzen badugu, a ren balioa era honetan lor dezakegu

$$a = \frac{B_y}{n} - b\frac{B_x}{n} = \bar{y} - b\bar{x}$$

Erregresio-zuzena $(\bar{x}, \bar{y})$ puntutik pasatzen da eta puntu hau diagramaren grabitate-zentroa da.

$$y = a + bx \Rightarrow y = \bar{y} - b\bar{x} + bx$$

beraz

$$y - \bar{y} = b(x - \bar{x})$$

Korrelazio koefizientea

Bi dimentsiotako laginen korrelazio linealaren koefizientea era honetan definitzen da,

$$r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{Kob(x,y)}{\sqrt{Barx}\sqrt{Bary}}$$

r koefizientearen balioak $[-1,1]$ tartean daude eta x eta y-en arteko dependentzia lineala neurtzen du, $|r|$-ren balioa 1-ari hurbiltzen denean dependentzia handia da eta zeinuak zuzenaren maldaren zeinua adierazten du.

Kurba esponentzialen doiketa

Aldagaien arteko erlazioa esponentziala denean $y = A\mathrm{e}^{bx}$ erako funtzioa dugu. Logaritmo nepertarrak harturik

$$\ln y = \ln A + bx$$

lortzen dugu eta, $u = \ln y$ eta $a = \ln A$ aldaketak egiten $u = a + bx$ zuzena lortzen dugu. Doiketa egiteko, erregresio linealaren metodoa erabiltzen dugu $(x_1, \ln y_1), (x_2, \ln y_2), \ldots, (x_n, \ln y_n)$ puntu-multzoekin.

Kurba potentzialen doiketa

Aldagaien arteko erlazioa potentziala denean $y = Ax^B$ erako funtzioa dugu. Berriro logaritmo nepertarrak hartuz

$$\ln y = \ln A + B \ln x$$

lortzen dugu, $v = \ln y$, $a = \ln A$ eta $u = \ln x$ aldaketak egiten

$$v = a + bu$$

zuzena lortzen dugu, eta parametroak lortzeko erregresio linealaren metodoa erabiltzen dugu $(\ln x_1, \ln y_1), (\ln x_2, \ln y_2), \ldots, (\ln x_n, \ln y_n)$ puntu-multzoekin.

Parabolen doiketa

x eta y aldagaien arteko erlazioa bigarren mailako polinomio baten bidez adierazten da.

$$y = a + bx + dx^2$$

Parametroak lortzeko,

$$F(a, b, c) = \sum_{i=1}^{n} (y_i - a - bx_i - dx_i^2)^2$$

funtzioaren minimoak bilatzen ditugu zuzenaren kasuan bezala parametroekiko deribatu partzialak zero eginaz.

Parabolen doiketa

$$\frac{\partial F}{\partial a} = 2\sum_{i=1}^{n}(y_i - a - bx_i - dx_i^2)(-1) = 0$$

$$\frac{\partial F}{\partial b} = 2\sum_{i=1}^{n}(y_i - a - bx_i - dx_i^2)(-x_i) = 0$$

$$\frac{\partial F}{\partial d} = 2\sum_{i=1}^{n}(y_i - a - bx_i - dx_i^2)(-x_i^2) = 0$$

Hiru ekuazio horietatik sistema hau lortzen da, eta ebatzi behar dugu

$$\begin{cases} a \cdot n & +b \cdot B_x & +d \cdot B_{xx} & = B_y \\ a \cdot B_x & +b \cdot B_{xx} & +d \cdot B_{xxx} & = B_{xy} \\ a \cdot B_{xx} & +b \cdot B_{xxx} & +d \cdot B_{xxxx} & = B_{xxy} \end{cases}$$