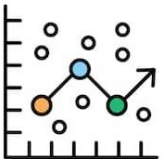


РЕГРЕСИЈА И КОРЕЛАЦИЈА

Приликом доношења пословних одлука често постоји потреба за предвиђањем вредности неког обележја на основу његове повезаности са једним или више других обележја.

Та међусобна повезаност нарочито је изражена код економских појава које су по својој природи динамичне и подложне променама. Тако је, на пример, познато да тражња за неким производом широке потрошње, између осталог, зависи од његове цене, цене конкурентних производа, дохотка по глави становника и сл. За валидну прогнозу неопходно је познавање природе зависности тражње од ових и других обележја која на њу утичу.

Под појмом регресиона и корелациона анализа подразумева се скуп статистичких процедура за испитивање степена и облика зависности између два или више обележја.

Основни смисао регресионе анализе јесте у откривању *постојања*, *облика* и *смера* везе између посматраних појава, док се корелациона анализа фокусира на испитивање *јачине*, односно *интензитета слагања* посматраних појава.

Анализа пословних и економских процеса се, у великој мери, заснива на испитивању односа између обележја. Овај однос се математички може приказати на следећи начин:

$$Y = f(X)$$

где функција може бити и у линеарном и у нелинеарном облику.

Регресиони модел којим изражавамо линеарну везу између две промењиве назива се линеарни регресиони модел, док се у осталим случајевима ради о нелинеарном регресионом моделу. Линеарне везе имају широку пословну и економску примену.

ПРОСТА ЛИНЕАРНА РЕГРЕСИЈА

Једначина регресије представља математички израз којим се описује природа посматране зависности. Ако се испитује зависност између два обележја, тј. ако у математичком изразу постоје само две промењиве, једна зависна и једна независна, и ако се оригиналним подацима ових обележја може прилагодити линеарна функција, одговарајућа регресија назива се проста.

У случају да промена независне променљиве (X) узрокује промену зависне променљиве (Y) у истом смеру, реч је о директној вези. Са друге стране, уколико порасту једне променљиве одговара опадање друге и обрнуто, реч је о инверзној вези.

У случају већег броја независних променљивих регресија је сложена или вишеструка. Даље ће бити речи о простој линеарној регресији.

Дијаграм распршености

Први корак у испитивању зависности између два обележја је графичко приказивање емпиријских података. Подаци се представљају у правоуглом координатном систему у равни, при чему сваком пару података за независну и зависну променљиву одговара тачка у координатној равни. Нека су (x_i,y_i), $i = 1,2,\ldots,n$ парови узорачких података за обележја X и Y чију зависност испитујемо. Овакав графикон се назива дијаграм растурања (распршености). На основу дијаграма растурања може се стећи визуелни утисак о постојању и природи зависности између посматраних обележја.

Пример 1.

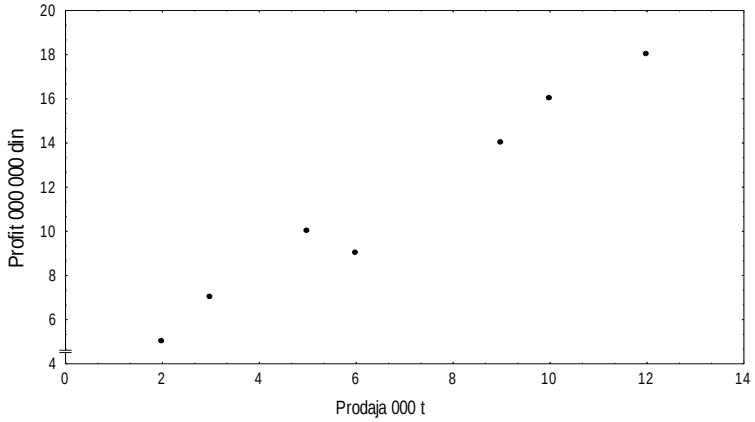
Нацртати дијаграм распршености на основу података о продаји и профиту једног предузећа. Подаци су приказани у следећој табели:

Табела 1.

Продаја (у 000 t)	Профит (у 000 000 дин.)
2	5
3	7
5	10
6	9
9	14
10	16
12	18

Решење:

Може се уочити да су тачке на дијаграму растурања распоређене приближно око праве линије, одакле се може наслутити да између профита и продаје постоји линеарна зависност. Што се тачке више групишу око замишљене праве линије, односно што су ближе овој линији, веза између променљивих је чвршћа и обрнуто. У случају да све тачке леже на правој, реч је о функционалној вези, што је изузетно редак случај у економији.



Слика 1. Дијаграм распршености

Метод најмањих квадрата

Прво и основно питање које се поставља код линеарне регресије јесте како на основу парова узорачких података (x_i,y_i), $i = 1,2,...,n$, одредити најбоље прилагођену линију, тј. како оценити једначину линеарне регресије.

Општи облик линеарне регресије је:

$$\hat{y}_i = a + b \cdot x_i \text{ , } i = 1,2,...,n$$

где су a и b оцене непознатих коефицијената у линеарној функцији.

Оцене непознатих коефицијената добијају се преко следећих образаца:

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^{\,2} - n\bar{x}^{\,2}}$$

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

где су $\bar{x}$ и $\bar{y}$ аритметичке средине узорачких опсервација за независну и зависну променљиву.

Пример 2.

На основу података о продатој количини робе и профиту предузећа из претходне табеле, оценити линеарну регресију и представити је графички заједно са узорачким опсервацијама на дијаграму распршености.

$\hat{y} = a + bx$

Решење:

Елементи за израчунавање оцена приказани су у наредној табели:

Табела 2.

Продаја 000 т x_i	Профит (у 000 000 дин). y_i	$x_i y_i$	x_i^2
2	5	10	4
3	7	21	9
5	10	50	25
6	9	54	36
9	14	126	81
10	16	160	100
12	18	216	144
47	79	637	399

Пошто је у овом примеру $n = 7$, онда је:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{47}{7} = 6{,}71 \quad ; \quad \bar{y} = \frac{\sum y_i}{n} = \frac{79}{7} = 11{,}29$$

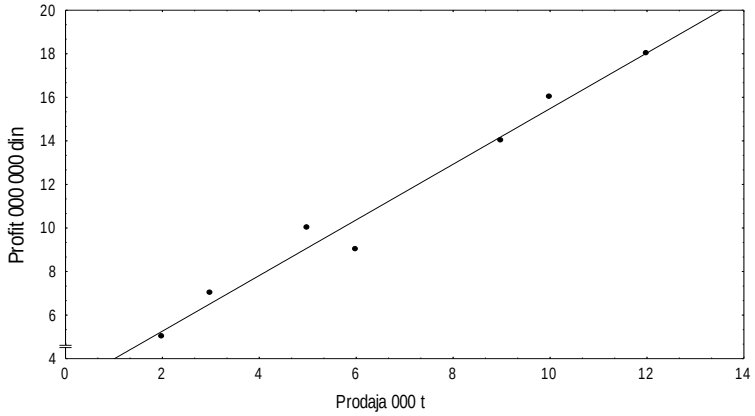
$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^{\,2} - n\bar{x}^2} = \frac{637 - 7 \cdot 6{,}71 \cdot 11{,}29}{399 - 7 \cdot 6{,}71^2} = \frac{106{,}71}{83{,}83} = 1{,}27$$

$$a = \bar{y} - b\bar{x} = 11{,}29 - 1{,}27 \cdot 6{,}71 = 2{,}77$$

Према томе, оцењена једначина регресије је:

$$\hat{y}_i = 2{,}77 + 1{,}27 \cdot x_i$$

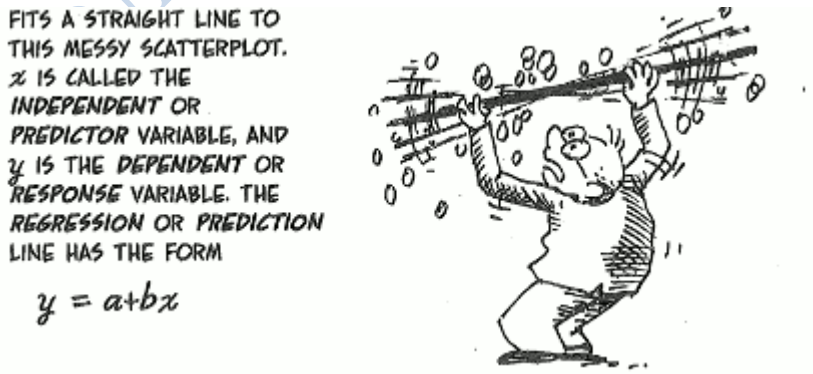
Како је права линија једнозначно одређена са две своје тачке, за њено графичко представљање довољно је да одредимо било које две тачке које јој припадају. На пример, за x_i=1, следи да је $\hat{y}_i = 2{,}77 + 1{,}27 \cdot 1 = 4{,}04$, а за x_i=13, следи да је $\hat{y}_i = 2{,}77 + 1{,}27 \cdot 13 = 19{,}28$. Дакле, тачке са координатама (1;4,04) и (13;19,28) припадају линији регресије. Њиховим спајањем у координатној равни добија се график линије регресије.



Слика 2. Оцењена линија регресије

Коефицијент a представља пресек са ординатном осом, односно одсечак на Y оси на дијаграму растурања. Ова оцењена вредност у пракси обично нема посебну важност, али како је то вредност зависне променљиве кад независна променљива има вредност нула, у нашем примеру се може рећи да у случају када нема улагања у рекламу профит износи 2.770.000 дин. Коефицијент b код праве линије представља коефицијент њеног правца, односно величину промене зависне променљиве када се независна променљива промени за јединицу. Стога се може закључити да када се улагања у рекламу повећају за 1.000 динара профит се у просеку повећава за 1.270.000 дин.

Предзнак који стоји уз коефицијент b указује на смер слагања појава, односно уколико је $b>0$ између посматраних променљивих постоји директна веза, док уколико је $b<0$, веза између променљивих је инверзна.



Стандардна грешка регресије

Стандардна грешка регресије је апсолутна мера варијације оригиналних (емпиријских) података зависне променљиве Y од оцењених вредности, као што је и стандардна девијација мера варијације сваке појединачне вредности од аритметичке средине. Другим речима, стандардна грешка је стандардна девијација око линије регресије.

Емпиријски подаци могу у мањој или већој мери одступати од линије регресије, с тим што уколико су та одступања мања, веза између зависне и независне променљиве је јача. Стандардна грешка регресије, као мера тих одступања, може се израчунати по формули:

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Да би стандардна грешка регресије могла да се одреди морају да се израчунају вредности $\hat{y}_i$ са линије регресије за све узорачке вредности независне променљиве x_i. У случајевима када постоји велики број таквих вредности израчунавање стандардне грешке може бити обиман посао. Међутим, стандардна грешка се може израчунати и без оцењених вредности $\hat{y}_i$ коришћењем оцењених коефицијената a и b, односно:

$$s_e = \sqrt{\frac{\sum y_i^{\,2} - a \cdot \sum y_i - b \cdot \sum x_i y_i}{n-2}}$$

С обзиром да представља просечно одступање узорачких опсервација од линије регресије, стандардна грешка регресије показује колико добро оцењена регресија описује зависност посматраних обележја. Када би стандардна грешка регресије била једнака нули, то би значило да оцењена регресија идеално оцењује зависну променљиву.

Пример 3.

У наредној табели приказани су подаци о броју кишних дана и броју ноћења у јулу месецу у последњих 10 година у једном хотелском комплексу на мору. На основу датих података:

а) Оценити једначину линеарне регресије и представити је графички;

б) Израчунати стандардну грешку регресије помоћу оцењених вредности са линије регресије.

Табела 3.

Број кишних дана	Број ноћења (у 000)
0	186
1	173
1	158
2	165
2	144
3	128
3	134
3	110
4	105
5	98

Решење:

а) Елементи за израчунавање оцена коефицијената приказани су у табели:

Табела 4.

Број кишних дана x_i	Број ноћења 000 y_i	$x_i y_i$	x_i^2
0	186	0	0
1	173	173	1
1	158	158	1
2	165	330	4
2	144	288	4
3	128	384	9
3	134	402	9
3	110	330	9
4	105	420	16
5	98	490	25
24	1401	2975	78

Пошто је у овом примеру $n = 10$ онда је:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{24}{10} = 2{,}4 \quad : \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1401}{10} = 140{,}1;$$

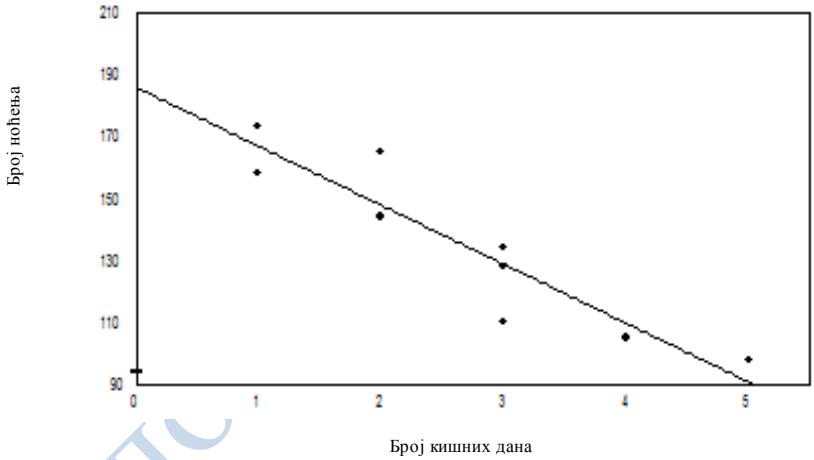
$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^{\,2} - n\bar{x}^2} = \frac{2975 - 10 \cdot 2{,}4 \cdot 140{,}1}{78 - 10 \cdot 2{,}4^2} = -\frac{387{,}4}{20.4} = -18{,}99$$

$$a = \bar{y} - b\bar{x} = 140{,}1 + 18{,}99 \cdot 2{,}4 = 185{,}68$$

Према томе, једначина регресије је:

$$\hat{y}_i = 185{,}68 - 18{,}99 \cdot x_i$$

Две тачке са ове праве су нпр. (0; 185,68) и (5; 90,73). Дијаграм растурања са линијом регресије приказан је на следећем графикону:



Слика 3. Оцењена линија регресије

У овом примеру коефицијент b је негативан што значи да је регресија опадајућа. Дакле, ако се број кишних дана повећа за један број ноћења се смањује у просеку за 18.990.

б) За стандардну грешку регресије прво треба одредити оцењене вредности са линије регресије. Оне се добијају тако што се узорачке вредности за независну промењљиву замењују у оцењену једначину регресије. Потребна израчунавања приказана су у следећој табели:

Табела 5.

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
186	185,68	0,32	0,1024
173	166,68	6,31	39,8161
158	166,68	-8,68	75,3424
165	147,70	17,3	299,2900
144	147,70	-3,7	13,6900
128	128,71	-0,71	0,5041
134	128,71	5,29	27,9841
110	128,71	-18,71	350,0641
105	109,72	-4,72	22,2784
98	90,73	7,27	52,8529
		$\approx 0{,}00$	881,9245

Стандардна грешка регресије је:

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{881{,}9245}{10-2}} = \sqrt{110{,}2406} = 10{,}50\,.$$

Др Наташа Папић-Благојевић, проф.