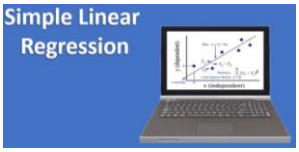


РЕГРЕСИЈА И КОРЕЛАЦИЈА

Предвиђање помоћу једначине регресије

Израчуната стандардна грешка регресије даје могућност да се, на основу утврђеног регресионог модела за вредност независне променљиве која се не налази у узорку x_0, предвиди одговарајућа вредност зависне променљиве y_0.

На основу једначине регресије може се израчунати вредност $\hat{y}_0 = a + b \cdot x_0$ која представља оцену стварне вредности y_0. Непозната вредност y_0 не може се тачно предвидети, али се може одредити интервал поузданости који је покрива са одређеном вероватноћом.

Под претпоставком да су узорачке опсервације независне и да имају нормалну расподелу са истом варијансом, може се показати да интервал:

$$\hat{y}_0 - t_{\alpha/2;n-2} \cdot s_p < y_0 < \hat{y}_0 + t_{\alpha/2;n-2} \cdot s_p$$

са вероватноћом $1-\alpha$ покрива вредност y_0, где је $t_{\alpha/2;n-2}$ квантил t - расподеле са $n-2$ степени слободе, а s_p стандардна грешка прогнозе која се израчунава по формули:

$$s_p = s_e \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^2 - n \cdot \bar{x}^2}}$$

Уколико су веће вредности стандардне грешке прогнозе, интервал поверења за дати ризик α биће шири, односно непрецизинији.

Вредности које су садржане у обрасцу за одређивање стандардне грешке прогнозе, утичу на њену величину. Тако, на пример, са повећањем вредности стандардне грешке регресије s_e, повећава се и стандардна грешка прогнозе s_p, док се са повећањем величине узорка n, стандардна грешка прогнозе s_p смањује.

Такође, уколико се вредност независне променљиве која се не налази у узорку x_0 удаљава од аритметичке средине $\bar{x}$, стандардна грешка прогнозе се повећава, а интервал се шири и тиме постаје непрецизнији.

Пример 1.

У наредној табели приказани су подаци о ценама и продатој количини тостер апарата једног произвођача:

Цена (у 000 дин/ком)	Продаја (у ком)
2,8	520
3,0	550
3,4	480
3,6	430
4,0	420
4,1	400
4,2	420
4,5	380

а) Оценити линеарну регресију и представити је графички;

б) Предвидети са 95% поузданости колико ће се продати тостер апарата по цени од 5.000 дин/ком.

Решење:

а) Потребни подаци за израчунавање оцена коефицијената и стандардне грешке регресије приказани су у следећој табели:

Табела 1.

Цена (у 000 дин/ком) x_i	Продаја (у ком) y_i	$x_i y_i$	x_i^2	y_i^2
2,8	520	1.456	7,84	270.400
3,0	550	1.650	9,00	302.500
3,4	480	1.632	11,56	230.400
3,6	430	1.548	12,96	184.900
4,0	420	1.680	16,00	176.400
4,1	400	1.640	16,81	160.000
4,2	420	1.764	17,64	176.400
4,5	380	1.710	20,25	144.400
29,6	3600	13.080	112,06	1.645.400

Како је у овом примеру n= 8 (наведени број података), следи да је:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{29,6}{8} = 3,7\,;\ \bar{y} = \frac{\sum y_i}{n} = \frac{3600}{8} = 450\,;$$

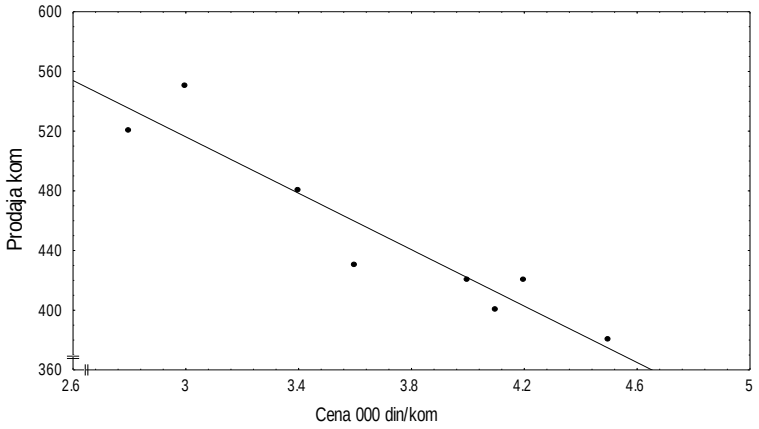
$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^{\,2} - n\bar{x}^2} = \frac{13080 - 8 \cdot 3,7 \cdot 450}{112,06 - 8 \cdot 3,7^2} = -\frac{240}{2,54} = -94,5$$

$$a = \bar{y} - b\bar{x} = 450 + 94,5 \cdot 3,7 = 799,65$$

Према томе, оцењена једначина регресије је:

$$\hat{y}_i = 799{,}65 - 94{,}5 \cdot x_i$$

Две тачке са ове праве су нпр. (2,5;564,4) и (4,5;374,4). Дијаграм распршености са линијом регресије приказан је на следећем графикону:



Слика 1. Оцењена линеарна регресија

б) Стандардна грешка регресије је:

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

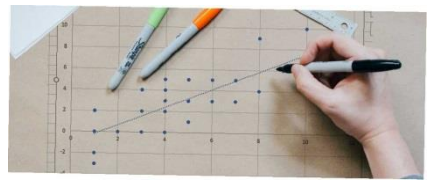
$$= \sqrt{\frac{1645400 - 799{,}65 \cdot 3600 + 94{,}5 \cdot 13080}{8-2}} = \sqrt{\frac{2720}{6}} = \sqrt{453{,}33} = 21{,}29$$

За $x_0 = 5$ је

$$\hat{y}_0 = 799{,}65 - 94{,}5 \cdot 5 = 327{,}15 \ .$$

Дакле, на основу једначине регресије, при цени од 5.000 дин., може се очекивати продаја од 327,15 односно ≈ 327 комада тостер апарата. Стандардна грешка прогнозе је:

$$s_p = s_e \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum x_i^{\,2} - n \cdot \bar{x}^2}} =$$



$$= 21{,}29 \cdot \sqrt{1 + \frac{1}{8} + \frac{(3{,}7-5)^2}{112{,}06 - 8 \cdot 3{,}7^2}} = 21{,}9 \cdot \sqrt{1{,}79} = 21{,}29 \cdot 1{,}34 = 28{,}53.$$

За 95%-ни интервал поузданости је α=0,05 па је $\alpha/2$=0,025. За $n-2=8-2=6$ степени слободе је $t_{0{,}975;6} = 2{,}45$.* Заменом ових вредности у формули:

$$\hat{y}_0 - t_{\alpha/2;n-2} \cdot s_p < y_0 < \hat{y}_0 + t_{\alpha/2;n-2} \cdot s_p$$

добија се да је

$$327{,}15 - 2{,}45 \cdot 28{,}53 < y_0 < 327{,}15 + 2{,}45 \cdot 28{,}53$$

тј.

$$257{,}25 < y_0 < 397{,}05.$$

Дакле, са 95% поузданости може се очекивати да ће при цени од 5.000 дин/ком бити продато између 257 и 397 тостер апарата.

* Објашњење:

За 95%-ни интервал поузданости је ризик грешке α = 0,05 јер је P (поузданост) = 95%=0,95, а $P + \alpha = 1$. Пошто нам је потребна таблична вредност за $\alpha/2$, следи да је $\alpha/2$=0,025. У таблицама се налази вредност за 1-$\alpha/2$ = 0,975 и за $n-2=8-2=6$ степени слободе, па је $t_{0{,}975;6} = 2{,}45$.

Студентова t дистрибуција (ниво поузданости 99% и 95%)

n	$t_{.995}$	$t_{.975}$
1	63.66	12.71
2	9.92	4.30
3	5.84	3.18
4	4.60	2.78
5	4.03	2.57
6	3.71	**2.45** ← таблична вредност
7	3.50	2.36
8	3.36	2.31
9	3.25	2.26
10	3.17	2.23
...	...	...

Коефицијент корелације

Коефицијент корелације је показатељ степена линеарне зависности између два обележја или, другачије речено, показатељ степена квантитативног слагања између посматраних променљивих. У основном скупу коефицијент корелације се обележава са ρ, а у узорку са r. Надаље ћемо користити ознаку r за означавање коефицијента корелације.

За узорачке опсервације $(x_i, y_i), i = 1,2,\cdots,n$ коефицијент корелације се израчунава по формули:

$$r = \frac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{(\sum x_i^{\,2} - n \cdot \bar{x}^2) \cdot (\sum y_i^{\,2} - n \cdot \bar{y}^2)}}$$

Вредност коефицијента корелације је број који се налази између -1 и 1, тј. $-1 \le r \le 1$ или $|r| \le 1$. У случају када је коефицијент корелације по апсолутној вредности једнак јединици, тј. $|r| = 1$, између посматраних обележја постоји *потпуна линеарна зависност*, па се све тачке на дијаграму распршености налазе на правој. Кад је коефицијент корелације једнак нули, тј. $r = 0$, *обележја су независна*. Одавде се може наслутити да што је коефицијент корелације по апсолутној вредности ближи јединици линеарна зависност између обележја је јача, а што је ближи нули зависност је слабија.

Као што се види, коефицијент корелације може бити позитиван или негативан број. Ако вредности за обележја истовремено расту или опадају, коефицијент корелације је већи од нуле. У том случају је корелација између појава директна или позитивна. Ако вредности за једно обележје расту а за друго опадају коефицијент корелације је мањи од нуле, па је веза између посматраних обележја инверзна. Предзнак испред оцењеног коефицијента b такође указује какав је предзнак испред коефицијента корелације, односно ако је коефицијент b позитиван и коефицијент корелације ће бити позитиван. Такође, уколико је коефицијент b негативан и коефицијент корелације ће бити негативан.

Пример 2.

На основу података о цени и потрошњи једног производа:

Цена x_i	Потрошња y_i
10	80
12	76
15	71
18	65
20	60
23	55
25	45

Испитати степен зависности између ова два обележја.

Решење:

Подаци потребни за израчунавање коефицијента корелације приказани су у наредној табели:

Табела 2.

Цена x_i	Потрошња y_i	$x_i y_i$	x_i^2	y_i^2
10	80	800	100	6.400
12	76	912	144	5.776
15	71	1.065	225	5.041
18	65	1.170	324	4.225
20	60	1.200	400	3.600
23	55	1.265	529	3.025
25	45	1.125	625	2.025
123	452	7.537	2.347	30.092

Како је у овом примеру n=7, то је:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{123}{7} = 17{,}571; \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{452}{7} = 64{,}571;$$

На основу израчунатих вредности сада можемо одредити и коефицијент корелације:

$$r = \frac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{(\sum x_i^2 - n \cdot \bar{x}^2) \cdot (\sum y_i^2 - n \cdot \bar{y}^2)}}$$

$$= \frac{7.537 - 7 \cdot 17{,}571 \cdot 64{,}571}{\sqrt{(2.347 - 7 \cdot 17{,}571^2) \cdot (30.092 - 7 \cdot 64{,}571^2)}} = \frac{-405{,}039}{\sqrt{168.371{,}560}} = -\frac{405{,}039}{410{,}331} = -0{,}987$$

На основу израчунатог коефицијента корелације можемо закључити да између посматраних променљивих постоји веома јака негативна корелациона веза.

Коефицијент детерминације

Квадрат коефицијента корелације r^2 назива се коефицијент детерминације и представља релативну меру варијације.

Коефицијент детерминације r^2 показује колико се добро зависна променљива објашњава независном променљивом, а изражен у процентима $r^2 \cdot 100\%$ показује са колико се процената зависна променљива објашњава утицајем независне променљиве. Када је оцењена једначина регресије, коефицијент детерминације се може израчунати и помоћу формуле:

$$r^2 = \frac{a \cdot \sum y_i + b \cdot \sum x_i y_i - n \cdot \bar{y}^2}{\sum y_i^{\,2} - n \cdot \bar{y}^2},$$

а одавде се може коренивањем добити коефицијент корелације $r = \pm\sqrt{r^2}$, с тим да се узима онај знак који има и оцењени коефицијент b.

Пример 3.

На основу података о ценама и продатој количини тостер апарата из првог примера, израчунати коефицијент корелације коришћењем оцењене једначине регресије.

Решење:

У посматраном примеру добијени су следећи резулатати:

$$n = 8; \quad \bar{y} = 450; \quad \sum y_i = 3.600; \quad \sum x_i y_i = 13.080; \quad \sum y_i^{\,2} = 1.645.400$$

$$\hat{y}_i = 799{,}65 - 94{,}5 \cdot x_i$$

па је :

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

$$= \frac{799{,}65 \cdot 3600 - 94{,}5 \cdot 13080 - 8 \cdot 450^2}{1645400 - 8 \cdot 450^2} = \frac{22680}{25400} = 0{,}893$$

На основу израчунатог коефицијента детерминације закључујемо да је 89,3% варијација у продатој количини тостер апарата одређено висином цене.

Помоћу израчунатог коефицијента детерминације може се одредити и коефицијент корелације, с тим што се за предзнак узима предзнак коефицијента b који је негативан, па је:

$$r = -\sqrt{0{,}893} = -0{,}94.$$

Закључујемо да између посматраних променљивих постоји јака негативна веза, односно како се повећава цена тостер апарата, тако опада њихова продаја.

Др Наташа Папић-Благојевић, проф.