

06 – Упражнение 2023/2024

Регресионен анализ. Линейна регресия

ас.г-р Костагин Костадинов

Същност на регресионния модел

Когато две променливи са в корелация помежду си (било то положителна или отрицателна), познаването на стойността на едната променлива позволява да **предскаже**, сравнително точно, стойността на другата. Това позволява да се правят полезни прогнози. Подобни модели са изключително важни в епидемиологията и клиничната медицина, а най-често използваният метод за тяхното създаване е **линейната регресия**.

Променливата, чиито стойности се предсказват в регресионния модел се нарича зависима (результативна), докато променливата, която се използва за предсказване и построяването на модела се нарича независима (факторна). В своята същност, регресионния анализ достига до "уравнение" (математическа функция) чрез което в аналитичен вид се представя връзката между определен набор от фактори и резултативното явление. Регресионното уравнение се представя като:

$$y = a + \beta \cdot x$$

Нека приемем за зависима променлива y – тегло на студентите от 2-ри курс, а за независима x – калорийния им прием. Чрез линейната регресия връзката между тези променливи може да се моделира, така че при известни стойности за калорийния прием, да предскажем теглото на студента. Регресионният модел следва формулата: $y = a + \beta \cdot x$, в която:

- Коефициентът a е **константа** (intercept).

Стойността коефициента показва, какво е теглото, когато калорийният прием е нула. С други думи, a посочва началната точка на взаимовръзката между тегло и калорийния прием и на практика, няма практическо значение в интерпретацията на модела, тя служи единствено за математическото моделиране.

- β е **регресионният коефициент**.

Това е число, което показва **силата и посоката** на връзката между предиктора (калорийния прием) и резултативната променлива (теглото). Математически, за установяване на стойността на y , трябва да умножи регресионния коефициент β с известната стойност калорийния прием x . Отрицателните стойности на β означават негативна (обратна) корелационна връзка, а позитивните стойности – позитивна (права).

Регресионно моделиране

Създаването на линеен регресионен модел протича в следните стъпки:

1. Избор на зависима и независима променлива
2. Изчисляване на регресионния коефициент β
 1. Изчисляване на сумата на стойностите на променливите x и y , както и сумата на всички произведения между x по y за всяко едно наблюдение;
 2. Сбор на повдигнатите на квадрат стойности на променливата x и умножение на получената сума с броя на наблюденията;
3. Изчисляване на константата a
 1. Изчисляване на средната стойност на променливата y ;
 2. Изчисляване на средната стойност на променливата x ;
4. Интерпретация на модела.

За да се представят етапите в тяхната последователност, се използва следния пример

Пример

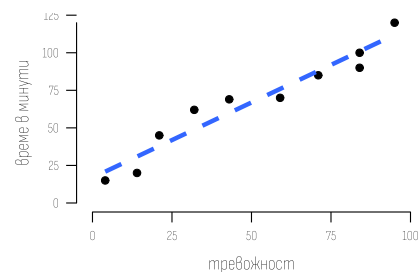
Проучване измерва степента на тревожност сред 10 студенти в специалност "медицина", посредством въпросник със скала от 0 (най-ниска степен на тревожност) до 100 (най-висока степен на тревожност). Изследователският екип записва резултата от теста, както и времето (в минути) прекарано в четене по статистика в предходния ден за всеки един участник. Целта на изследването е да се построи **линеен регресионен модел** за предсказване на тревожността на студентите y въз основа на времето прекарано в четене x .

Избор на зависима и независима променлива

Изходните данни са представени в таблица 1 и на фигура 1. Променливата тревожност е зависима, това означава, че изградения модел ще се опитва да *предскаже* стойността на тревожността във основа на променливата учене по статистика, която е независима или факторна.

таблица 1: Изходни данни за регресионен модел

y Тревожност (0 до 100)	x учене по статистика (минути)
21	45
84	90
14	20
95	120
84	100
32	62
59	70
43	69
71	85
4	15



Изчисляване на регресионния коефициент бета

За изчисление на този коефициент се ползва формулата:

$$\beta = \frac{(n \cdot \sum (X \cdot Y)) - (\sum X \cdot \sum Y)}{(n \cdot \sum X^2) - (\sum X)^2}$$

Изчисляване на сумата на стойностите на променливите x и y , както и сумата на всички произведения между на x по y за всяко едно наблюдение;

В тази стъпка е необходимо да установи сборът на всички наблюдения за двете променливи. След това е необходимо да се установи сумата $\sum (X \cdot Y)$ от произведенията на стойностите на променливите x по y - за всеки един студент. Резултатите са представени таблично в таблица 2.

фигура 1: Връзка между тревожност и време за учене по статистика

В числител:

- n е броят на наблюденията;
- $\sum X$ е сумата от стойностите на независимата променлива x ;
- $\sum Y$ е сумата от стойностите на зависимата променлива y ;
- $\sum (X \cdot Y)$ е сумата от всички произведения на стойностите на x по стойностите на y за всяко едно наблюдение.

В знаменателя:

- n е броят на наблюденията;
- $\sum X^2$ е сумата от квадратите на стойностите на независимата променлива x ;
- $(\sum X)^2$ е квадратът на сумата от стойностите на независимата променлива x .

таблица 2: Първа стъпка за изчисляване на регресионния коефициент

y Тревожност	x учене	$x \cdot y$
21	45	945
84	90	7560
14	20	280
95	120	11400
84	100	8400
32	62	1984
59	70	4130
43	69	2967
71	85	6035
4	15	60
$\sum Y = 507$	$\sum X = 676$	$\sum (X \cdot Y) = 43761$

В резултат от тази стъпка се установява числителят в формулата за изчисляване на регресионния коефициент β

$$(n \cdot \sum (X \cdot Y)) - (\sum X \cdot \sum Y) = 10 \cdot 43761 - 676 \cdot 507 = 437610 - 342732 = 94878$$

Сбор на квадратите за променливата x

В тази стъпка се повдигат на квадрат всички стойности на променливата x , след което се сумират. Резултатът е представен в таблица 3.

таблица 3: Втора стъпка за изчисляване на регресионния коефициент

x учене по статистика (минути)	x^2
45	2025
90	8100
20	400
120	14400
100	10000
62	3844
70	4900
69	4761
85	7225
15	225
$\sum x = 676$	$\sum x^2 = 55880$

В края на втора стъпка получаваме знаменателя в формулата за изчисляване на регресионния коефициент β

$$(n \cdot \sum X^2) - (\sum X)^2 = 10 \cdot 55880 - 676^2 = 558800 - 456976 = 101824$$

Изчисляване на регресионния коефициент β

След като сме установили числителя и знаменателя замествахме във формулата:

$$\beta = \frac{(n \cdot \sum X \cdot Y) - \sum X \cdot \sum Y}{n \cdot \sum X^2 - (\sum X)^2}$$

$$\beta = \frac{94878}{101824} = 0,93$$

Изчисляване на константата a

За изчисляване на константата използваме формулата:

$$a = \bar{y} - (\beta \cdot \bar{x})$$

Където:

- $\bar{y}$ е средната стойност на зависимата променлива y ;
- $\bar{x}$ е средната стойност на независимата променлива x .
- β е регресионният коефициент.
- a е константата.

Изчисляване на средната стойност на зависимата променлива y

За да изчислим средната стойност на зависимата променлива y ползваме формулата за средна аритметична:

$$\bar{y} = \frac{\sum Y}{n}$$

$$\bar{y} = \frac{507}{10} = 50,7$$

Изчисляване на средната стойност на независимата променлива x

Отново прилагаме формулата:

$$\bar{x} = \frac{\sum X}{n}$$

$$\bar{x} = \frac{676}{10} = 67,6$$

Изчисляване на константата a

Заместваме в израза за константата:

$$a = \bar{y} - (\beta \cdot \bar{x})$$

$$a = 50,7 - (0,93 \cdot 67,6)$$

$$a = -12,2$$

Интерпретация:

1. **Регресионният коефициент** е позитивно число – това означава, че увеличаването на времето в четене по статистика, се свързва с увеличение на тревожността. Връзката е сравнително силна, като увеличение с една минута четене води до увеличение с 0,93 точки в теста за тревожност;
2. **Константата “a”** има значение единствено като “начална точка”. Това число показва стойността на зависимата променлива (тревожност), ако фактора (време прекарано в четене) е равен на 0.

Математически моделът се представя като:

$$y = a + \beta \cdot x = -12,2 + 0,93 \cdot x$$

Регресионно прогнозиране (използване на модела за предсказване)

Да приемем, че нов студент се премести в групата. Знаейки само колко време чете по статистика е необходимо да преценим, какъв би бил резултатът му от теста за тревожност. Да приемем, че студентът е споделил, че прекарва в четене 30 мин на ден. Прилагаме модела ¹:

$$y = a + \beta \cdot x = -12,2 + 0,93 \cdot x$$

За новия студент $x = 30$ минути. Заместваме в уравнението:

$$y = -12,2 + 0,93 \cdot 30 = 16$$

Оценка на регресионния модел

Всеки регресионен модел се оценява спрямо **способността му да предсказва правилно**. Тази оценка се извършва чрез показателя коефициент на определеност (**детерминирания**) – R^2 . Този коефициент:

1. Има стойности между 0 и 1 (0 и 100%);
2. Определя силата на модела – способността му да предсказва правилно;
3. Показва каква част от вариацията на зависимата променлива се обяснява от вариацията на независимата променлива;
4. Позволява да се оцени коефициентът на корелация, чрез формулата:

$$|r| = \sqrt{R^2}$$

В посоченият по-горе пример, ако приемем, че коефициентът на детерминирания е 0,93, това означава, че **93%** от вариацията в тревожността в групата се обяснява следствие на вариацията в времето за четене по статистика. Двете променливи са в права (позитивна) корелационна връзка, чиито коефициент е:

$$|r| = \sqrt{R^2} = \sqrt{0,93} = \pm 0,946$$

¹ Много компютърни програми използват методи подобни на регресията за целите на маркетинга. Някои интернет страници запамятват последните ви търсения и в основа на данни за вас (независими променливи или фактори) ви “предлагат” определени продукти, които за които има по-голяма вероятност да се интересувате (зависима променлива). В медицината регресия се използва наравно с медицинската слушалка. В кардиологията например рискът от инсулт се изчислява, използвайки именно регресионен модел наречен CHADVAS score

Общо правило за интерпретация на регресията:

1. Регресионният коефициент β , показва, как се променя резултатът (зависимата променлива) при **една единица повишаване на фактора** (предиктора).
2. Позитивен регресионен коефициент означава позитивна корелационна връзка между фактора и резултата. Обратно, негативен регресионен коефициент означава обратна (негативна) връзка между фактора и резултата.
3. R^2 представя какъв дял от вариацията на резултативната променлива се дължи на изследвания фактор.
4. Когато коренуваме стойността на R^2 получаваме **абсолютната стойност на корелационния коефициент** $\sqrt{R^2} = |r|$.
5. Знакът + или - пред коефициентът на корелация **се определя** от знака пред регресионния коефициент β .

Още малко за регресията

i Множествена регресия

Представения до момента регресионен модел съдържа само една "обясняваща" променлива. Затова такива модели се означават като "еднофакторни". В действителност, за да обясним една променлива, често се използват множество предиктори (независими променливи). Всеки един от тях има собствен регресионен коефициент. В тези случаи често се наблюдава статистическия феномен "колинеарност". При него две или повече независими са в **силна корелационна връзка помежду си**, при което тяхната способност да предскажат зависима променлива намалява.