06 - Упражнение

Регресионен анализ. Линейна регресия

ас.д-р Костадин Костадинов

Същност на регресионния модел

Корелацията разглежда връзката между две количествени променливи. Статистическата връзка се ползва, като едно от доказателствата за причинно-следствена връзка, но сама по себе си една статистическа връзка не е причинно следствена.

Регресионния анализ се базира на корелационния. Затова може да се твърди, че всички съображения за корелацията важат и тук. Установяването на връзка между променливите не означава, че тя е причинно следствена. Въпреки това, за разлика от корелацията при регресията се опитваме да създадем модел. Регресионния модел приема, че една от променливите е независима (нарича се още фактор, предиктор или "причина"), която действа върху зависимата променлива (нарича още резултат или резултативна променлива). При регресионния модел се изгражда "уравнение" с което можем да установим, точно колко е зависимостта на независимата върху зависимата променлива.

Например, като зависима променлива можем да използваме "тегло". При регресията, тази зависима променлива е "феноменът" който се опитваме да обясним и се бележи с y. Независима променлива ще изберем например калорийния прием. Независимите фактори, са тези чрез който се опитваме да обясним зависимата. Независимия фактор се означава с x. Нека приемем, че сме установили корелация (връзка) между тези две променливи. С други думи, увеличаването на

калорийния прием е свързано с увеличаване на теглото. При регресията искаме да моделираме тази връзка [1]. Приемаме, че “теглото” зависи от “калорийния прием”. Чрез регресионния модел, целим да опишем точно колко е тази зависимост, каква е нейната посока и сила, дали тя е значима (статистически) и дали моделът е подходящ да опише тази връзка. Математически, крайния модел излежда така

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

Регресионно моделиране (създаване на модела)

Нека опитаме да направим подобен регресионен модел. В случая ще използваме една студентка група и нивото на тревожност (измерено на скала от 0 до 100) като зависима променлива. За независима променлива ще използваме времето прекарано в учене по статистика (в минути).

Table 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

`geom_smooth()` using formula 'y ~ x'

Както прави впечатление от таблицата, а и от Figure 1, опредено такава връзка съществува. Но за да се моделира връзката следва да се изпозлваме следните стъпки:

[1] В нашият пример y е теглото - променливата която искаме да обядним. x е калорийния прием, чрез който искаме да обясним теглото. Коефициентът a се нарича константа (intercept). Стойността на този коефициент ни показва, какво е теглото, когато калорийния прием е нула. С други думи a ни посочва началната точка на връзката между тегло и калорийния прием, която сме установили. Целта на константа е да посочи откъде “започва” наблюдаваната връзка, тоест каква е стойността на теглото, при липса на стойност за калорийния прием. На практика, контанстата няма практически интерпретация, а само математична. β се нарича регресионния коефициент. Това е число, по което трябва да умножим калорийния прием за да установим теглото. Този коефициент показва силата и посоката на връзката. Голяма стойност за този коефициент означава силна връзка. Дори 1 калория ще доведе до много голяма стойност за теглото. Ниска стойност за този коефициент означава слаба връзка. Регресионния коефициент бета може да бъде и отрицателно число. Това означава, че връзката е негативна (обратна) с увеличаване на независимата променлива, зависимата ще намалява.

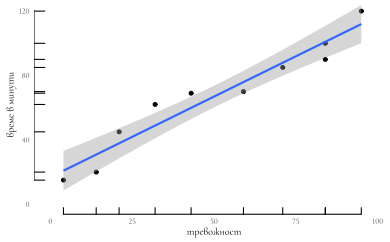

Figure 1: Връзка между тревожност и време прекарано в учене по статистика

Изчисляване на регресиония коефициент бета

За изчисление на този коефициент се ползва формулата:

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

Първо да погледнем числителя:

$$(n \cdot \sum (X \cdot Y)) - (\sum X \cdot \sum Y)$$

Необходимо е да умножим стойността по x със стойността по y за всяко наблюдение. След това сумираме всички тези произведения и ги умножаваме по броя на наблюденията. От получената стойност ще се извади произведението на сбора на всички стойности за x и всички стойности по y. Това представено таблично -

Table 2: Първа стъпка

y Тревожност (0 до 100)	x учене по статистика (минути)	$x \cdot y$
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
$\sum Y = 507$	$\sum X = 676$	$\sum X.Y = 43761$

След тази първа стъпка можем да установим колко е числителя:

$$(n \cdot \sum X \cdot Y) - (\sum X \cdot \sum Y) =$$

$$= 10 \cdot 43761 - 676 \cdot 507 = 437610 - 342732 = 94878$$

Сега вече можем да преминем към знаменателя

$$(n \cdot \sum X^2)) - (\sum X)^2$$

В знаменателя се интересуваме само от независимата променлива - в нашия случай това е времето учене по статистика. Необходима ни е сумата от всички времена повдигната на квадрат умножена по броя на участниците, както и сумата от квадратите на всички времена. Нека представим това таблично:

Table 3: Втора стъпка

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
$\sum x = 676$	$\sum x^2 = 55880$

Сега вече можем да заместим във формулата и да разберем колко всъщност е регресионния коефициент в случая

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

$$\beta = \frac{94878}{10 \cdot 55880 - 676^2} = \frac{94878}{101824}$$

$$\beta = 0,9317$$

Интерпретация:

Регресионният коефициент е позитивно число - това означава, че увеличаването на минутите прекарани в четене по статистика, се увеличава тревожността. Също така регресионният коефициент ни показва каква е силата на тази връзка. В случая една минута прекарана в четене на този документ носи 0,93 точки повече тревожност на четящия[2].

[2] Статистически установяваме каква е силата (тежестта) на фактора върху резултата, както и каква е посоката на действието му. Този коефициент е изчислен на базата на вече известни тревожност и прекарано време в четене. Той служи за описание на действието на фактора, върху резултата точно в тази студенска група.

Изчисляване на константата

За изчисляване на константата използваме формулата:

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

Необходимо ни е да се намерят средните стойности на предиктора (независимата променлива) $\bar{x}$ тоест средното време на учене в групата, както и средните стойности на зависимата променлива $\bar{y}$ тоест средното ниво на тревожност в групата.

Table 4: Определяна на средни стойности

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
$\bar{y} = 50,7$	$\bar{x} = 67,6$

Сега сме готови да заместим във формулата

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

$$a = 50,7 - (0,93 \cdot 67,6)$$
$$a = -12.2$$

Интерпретация на константата а

Константата “а” има значение единствено като “начална точка”. Тя ни показва колко би била стойността на зависимата променлива, ако фактора е равен на 0.

След установяването на константата и регресионния коефициент бета можем да запишем модела:

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

Регресионно прогнозиране (използване на модела за предсказване)

Да приемем, че нов студент се премести в групата. Знаейки само и единствено, колко време чете по статистика е необходимо да преценим, колко ще е тревожен. Да приемем, че студентът е споделил, че прекарва в четене 30 мин на ден. Можем да определим, колко е неговата тревожност в основа на модела.

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

Знаем, че този нов студент, чете статистика по 30 мин, тоест за него $x = 30$, заместваме в вече изградения модел и получаваме

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

Като използваме изградения модел за новия студент, имайки предвид че знаем, колко време чете по статистика, можем да предвидим нивото на тревожност. По този начин можем да използваме регресионния модел за да намали несигурността в бъдещето.[3]

[3] Много софтуерни платформи използват методи подобни на регресията за реклама. Някой от сайтовете търговските сайтове запаметяват последните ви търсения и в основа на данни за вас (използвани като независими променливи или фактори) ви "предлагат" нови продукти, които да купите (зависима променлива). В медицината използваме регресия наравно с медицинската слушалка. В кардиологията например риска от инсулт се изчислява използвайки именно регресионен модел наречен CHADVAS score

Оценка на регресионния модел

Колко е добър регресионния модел? Отговорът е толкова, колкото и модела на сграда преди да бъде построена. Понякога модела греши малко или много. За щастие ние имаме възможност да оценим тази грешка. С други думи, да установим статистическата сила на модела. Да се върнем обратно на примера с новия студент в групата. За него преди да установим реално, колко е тревожността му, имаме общо взето две възможности: да предположим, че неговата стойност ще е средната за група - 51 точки или да се доверим на модела и да предположим, че неговата стойност ще е 16. Каква е грешката от двете предположения можем да разберем единствено, ако студента направи теста за тревожност. Ако истинския резултат е по-близко до средната на групата това от нашия модел няма никаква полза. В другият случай, ако истинския му резултат е по-близо предсказания от модела - определено си заслужава да ползваме регресионния модел. Съществува начин да определим силата на статистическия модел. Това е индикаторът R^2. Нарича се още коефициент на **детерминирания.** Този индикатор установява колко е добър моделът ни. Неговата максимална стойност е 1-ца (или 100%). Този показател всъщност изразява колко по-добре нашия модел предсказва стойността на тревожността по-добре от средната аритметична за групата.

В нашият пример стойността на $R^2 = 0,93$. Интерпретира се като: 93% от вариацията в тревожността в групата се обяснява следствие на вариацията в времето за четене по статистика. Също така, можем да установим каква е корелацията между тревожността и времето прекарано в четене по статистика. Необходимо е да коренуваме коефициентът на детерминирания.

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

Дали да изберем знака + или - пред коефициентът на корелация, зависи от посоката на връзката: понеже от графиката, установяваме че с увеличаването на едната променлива,се увеличава и другата избираме знака +. С други думи в нашия пример има силна позитивна корелация между тревожността и времето прекарано в четене по статистика.

Общо правило за интерпретация на регресията:

Регресионният коефициент β, показва, как се променя резултатът (зависимата променлива) с една единица повишаване на фактора (предиктора). Позитивен регресионен коефициент означава позитивна връзка между фактора и резултата. Негативен регресионен коефициент означава обратна (негативна) връзка между фактора и резултата. R^2 установява, какъв процент от вариацията на резултативната променлива се дължи на факторната променлива. Когато коренуваме този коефициент получаваме абсолютната стойност на корелационния коефициент $\sqrt{R^2} = |r|$. [4]

[4] Понеже корен от което и да е число е винаги положително число, след като сме установили стойността на корелационния коефициент е необходимо да решим дали да сложим знак + или -. Това правим графично (като установим каква е посоката на връзката).

Още малко за регресията

Значимост на коефициентите

Всички коефициенти, които установихме, в крайна сметка са изчислени в основа на данните от извадката. Дали тези коефициенти (бета, константата a, корелационния коефициент) са значимо по-различни от 0-ла в генералната е съвкупност е въпрос на тестване на хипотези. Можем да приложим t тест и да установим дали например, коефициентът β е различен от нула в целия курс, а не само в една група. За да тестваме това, нулевата хипотеза би била, че този коефициент β, за генералната съвкупност е нула. При вероятност, да сме получили този коефициент случайно,

по-ниска от 0,05 отхвърляме нулевата хипотеза, приемаме, че този коефициент, в генералната съвкупност е значимо по-различен от 0-ла.

Множествена регресия

Обяснения до момента регресионен модел е опростен линеен модел. В действителност за да обясним една променлива не използваме само един фактор. Използваме множество предиктори. Всеки един от тях получава регресионен коефициент, формулата става доста по-дълга. Когато използваме повече от един предиктор се наблюдава интересен феномен - колинеарност това представлява взаимната зависимост между два или повече фактори в регресионния модел.

Нелинейни модели

Освен линейни модели, в статистиката са разработени още множество нелинейни модели. Всеки един от тях има различна формула, различна интерпретация на коефициентите, но общ смисъл. Търси се чисто математическа функция, чрез която да изградим уравнение. С помощта на това уравнение (модел) ще се опише връзката между фактор и резултат и ще може да се направи “предсказание” за стойноста на “новия неизвестен участник”.