14 ЛИНЕЙНА РЕГРЕСИЯ – ОЦЕНЯВАНЕ И ПРЕДВИЖДАНЕ

В глава 5 беше въведана концепцията за корелация между две случайни променливи. Беше показано, например, че когато корелацията е умерена и положителна, може да се очаква, че измерванията на едната променлива, които са по-големи от средните е вероятно да са свързани с измервания за другата променлива, които също са по-големи от съответната средна стойност. Тази глава фокусира върху използването на концепцията за корелацията и правата линия при въвеждането на друга концепция – тази за линейната регресия. Тази концепция предполага предвижданото[1] или оценяването на измерванията на едната променлива, на базата на знанието или наблюдаването на измерванията на другата променлива. В допълнение също ще може да се прави вероятностно заключение относно точността на прогнозата.

Предвижданото статистически смисъл е процесът на оценяване на измерванията на една променлива като се знаят стойностите на друга променлива, свързана по определен начин с първата.

Ключови термини	Предвиждане	Метод на най-малките квадрати
	Корелация	Грешка при предвиждането
	Регресионна линия	Стандартна грешка на оценката
	Точност на предвиждането	Хомоскедастичност
	Уравнение за предвиждане	Регресионен коефициент
	Предвидени стойности	Ковариация
		Доверителен интервал

Принципи на предвиждането

Да предположим, че група от 40 артефакта има средна стойност на дебелината от 8.8 мм. Ако искаме да получим оценка за един отделен артефакт, без да привличаме допълнителна информация, най-добрата оценка би била средната за цялата група, т.е. 8.8 мм. Ако допълнително е известно, че артефактът има ширина над средната и че има умерена положителна взаимовръзка между ширината и дебелината на артефактите, то може да се очаква, че индивидуалният резултат на конкретния артефакт ще е по-голям от 8.8 мм. Нещо повече, разликата между наблюдаваната индивидуална стойност и предвидената вероятно ще бъде по-малка от тази между индивидуалната стойност и средната от 8.8 мм. С други думи, когато предвиждаме стойността на една променлива, като използваме информация от свързани с нея други променливи, наблюдава се тенденция това предвиждане да е по-точно, отколкото предвиждането, което не използва знанията за други свързани променливи. Следователно, ние сме в състояние

[1] Ще правим разлика между термините "предвиждане" и "прогнозиране". Първият е свързан с понятието "интерполация", а вторият – с понятието "екстраполация".

да предвидим индивидуалния резултат на дебелината по-точно, когато имаме знания относно ширината, отколкото, когато тази информация отсъства.

В глава 5 беше въведена концепцията за линейна взаимовръзка между две променливи. Беше установено, че ако построим диаграмата на разсейване на две умерено корелиращи променливи, съответните двойки стойности ще лежат близко около една права линия (Фигура 5.2 а, б). Ако разгледаме всички двойки измервания за двете променливи, тогава тази права линия дава представа за средната промяна на стойностите на едната променлива при промяната на стойностите на другата. Тази линия се нарича *регресионна линия*. Ако използваме стойностите на променливата ***x***, за да предвиждаме стойностите на променливата ***y***, то тази линия се нарича регресионна линия на ***y*** върху ***x***. Често променливата ***x*** се нарича *независима* променлива, променливата ***y*** – *зависима.*

Линейната регресия е инструмент, който служи да се направи възможно най-точното предвиждане и е често най-приложимата процедура в изследователските условия на науки като археологията. Сега ще се обърнем към въпроса: какво означава да е налице точно предвиждане? Очевидно, необходимо е този термин да получи определение от статистическа гледна точка, а не просто логическо обяснение. Как да получим правата линия? Ние знаем, че тази линия трябва де бъде права, тъй като тук говорим за линейна регресия. Но всяка диаграма на разсейването, постороена за повече от две точки показва, че няма еднозначно решение, т.е. съществува множество прави линии, всяка от които минава през множеството точки. Коя линия да изберем? Отговорите на тези въпроси следват по-долу в контекста на един специфичен пример

Нека предположим, че имаме 20 артефакта, измерени по две променливи. Първата е ширината (***x***), а втората е дебелината (***y***). Ако ширината и дебелината на артефактите са тясно свързани, то може да се очаква забележимо висока и положителна корелация между двете множества стойности. С други думи, артефактите, които са по-голяма ширина е по-вероятно да имат и по-голяма дебелина. Данните за този пример са показани в Таблица 1, а съответната диаграма на разсейване – на Фигура 1.

Един поглед върху диаграмата на разсейването потвърждава нашите очаквания относно двойките стойности. По-високите резултати за ширината показват тенденция да са свързани с по-високи резултати за дебелината и, обратно, по-ниските резултати за ширината отговарят на по-ниски резултати за дебелината. Следователно, налице е визуално доказателство за естеството на връзката между тези две променливи, което дава основание да твърдим, че ще имаме положителен корелационен коефициент. Прилагайки формула (5.4) към тези данни получаваме за коефициента на корелацията стойността +0.74.

Стохастична прогностична връзка между две променливи

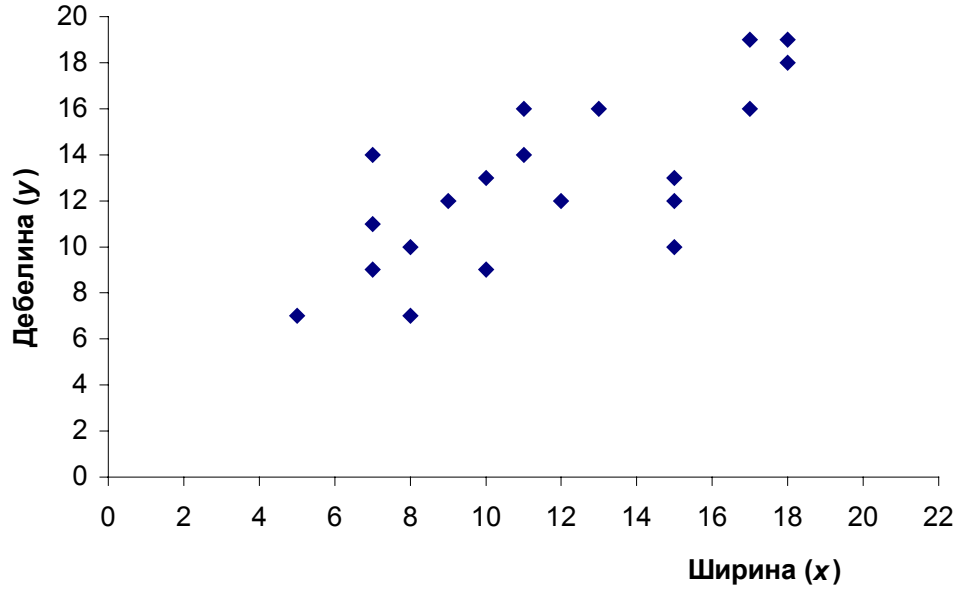
Обикновено между явленията в природата и в обществото съществуват два вида зависимости: а) функционални и б) статистически или стохастични.

В първия случай имаме еднозначно съответствие между явлението причина и явлението следствие. По друг начин са поставени нещата във втория случай.

Таблица 1. Данни за изчисляване на регресионното уравнение за предвиждане на дебелината на артефактите (y) на базата на тяхната ширина (x)

Артефакт	x	y	x^2	y^2	xy
1	15	12	225	144	180
2	10	13	100	169	130
3	7	9	49	81	63
4	18	18	324	324	324
5	5	7	25	49	35
6	10	9	100	81	90
7	7	14	49	196	98
8	17	16	289	256	272
9	15	10	225	100	150
10	9	12	81	144	108
11	8	7	64	49	56
12	15	13	225	169	195
13	11	14	121	196	154
14	17	19	289	361	323
15	8	10	64	100	80
16	11	16	121	256	176
17	12	12	144	144	144
18	13	16	169	256	208
19	18	19	324	361	342
20	7	11	49	121	77
Σ	**233**	**257**	**3037**	**3557**	**3205**
Средни:	$\bar{x}$**=11.65**	$\bar{y}$**=12.85**			

Фигура 1. Диаграма на разсейването за ширината и дебелината на артефактите



Статистическите закономерности могат да се проявят само при голям брой наблюдавани единици, в масов процес. При този тип закономерности на зададена стойност на зависимата променлива отговаря цяла редица стойности на обясняващата или независимата променлива, случайно разсеяни върху някакъв интервал. На всяка фиксирана стойност на независимата променлива съответства определена функция на разпределение на зависимата. Това е обусловено от факта, че зависимата променлива е подложена на влиянието на множество неконтролируеми или неотчетени фактори. Тъй като стойностите на зависимата променлива са случайно разсеяни, то не е възможно те да се предскажат точно, а само се оценявdt с определена вероятност.

От горното е ясно, че регресията може да се определи като едностранна стохастична зависимост, която установява съответствие между случайните променливи. Това съответствие обикновено се изразява във вид на функция, която се нарича *функция на регресията* или просто *регресия*. Ще отбележим, че за разлика от функционалната зависимост регресионната не е обратима. За това ще стане дума по-нататък

В науката, за да се обяснят получените от изследването резултати и да се разработи съответната теория, обикновено се използва т.нар. *принцип на икономичността*, който утвърждава, че най-полезно е най-простото обяснение. Тук, най-простото обяснение на връзката между променливите е линейността на тази връзка. Това означава, че математическото уравнение, което най-добре описва конфигурацията от диаграмата на разсейването е уравнението на правата линия. Следователно, нужно е да се определи коя права линия най-добре описва (възпроизвежда) взаимовръзката между двойките съответни точки.

Накратко, математическото уравнение на правата линия отразявя функционалната връзка между две променливи x и y. y трябва да е функция на x или обратното. Да предположим, че y е функция на x такава, че $y = bx + a$. Този вид на зависимостта се нарича *линейно уравнение с ъглов коефициент*. Тогава, ако a и b са известни лесно може да се определи стойността за y за всяка зададена стойност на x. На Фигура 2 е изобразена правата, зададена с уравнението си $y = 3x + 2$. Ако x=2, то y = 8. Ако x = -1, то y = -1 и т.н.

При линейната регресия линейното уравнение с ъглов коефициент има формата:

(1) $\hat{y} = bx + a$

където: $\hat{y}$ = предвидената стойност

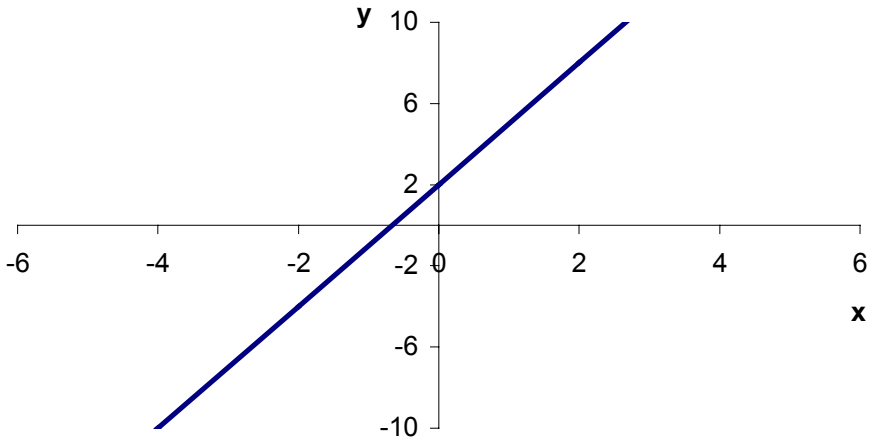
b = ъглов коефициент (коефициент на наклона, регресионен коефициент)

a = свободен член (точката, в която правата пресича ординатната ос y)

От Фигура 2 се вижда, че линията $y = 3x + 2$ пресича оста y в точката y = 2. Ъгловият (регересионният) коефициент b се определя съдържателно като промяната на y при единица промяна в x. В този пример, когато x нараства с единица y нараства с три единици. Затова регресионният коефициент е 3. При линейната регресия, определянето на уравнението на правата линия за дадено множество от съответни двойки измервания означава да се изчислят стойностите на коефициентите a и b.

За да се използва линейната регресия за предвиждане трябва да се построи права линия. Уравнението на правата линия е регресионното уравнение, използвано за предвиждане. Това уравнение има формата: $\hat{y} = bx + a$.

Фигура 2. График на $y = 3x + 2$

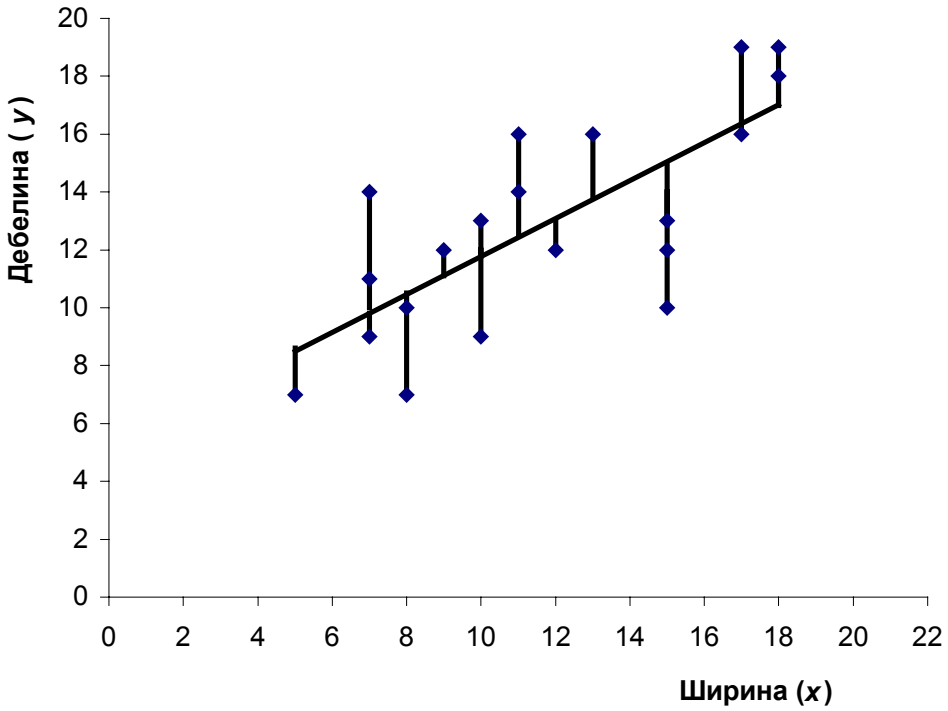


Определяне на регресионната линия

Вече е известна общата форма на уравнението на правата линия, но се появява въпросът коя линия от всички мъзможни трябва да бъде използвана в качеството на регресионна линия? Да предположим, че сме получили регресионното уравнение за предвиждане на дебелината на (***y***) по резултата за ширината (***x***), т.е. регресията на ***y*** върху ***x***. Ако се наблюдава идеална връзка между двете множества измервания, то всички точки от диаграмата на разсейването ще лежат върху права линия и уравнението на тази линия може да се определи лесно. Когато взаимовръзката между двете множества от точки не е идеална, изравняването на линията (определянето на ***a*** и ***b***) вече не е толкова лесна задача. Линията, която искаме да получим описва тенденцията или тренда в данните така, че промяната в променливата ***x*** ще се отрази като средна промяна в променливата ***y***. Изравняването на данните, за да се получи уравнението на правата (с други думи, оценяването на коефициентите ***a*** и ***b***) обикновено се извършва по *метода на най-малките квадрати (МНК)*. В нашия пример за предвиждането на дебелината (***y***) по ширината на артефакта (***x***) уравнението на линията се строи така, че сумата от квадратите на разстоянията от всички измерени точки до тази линия да е минимална. Най-добрата линия, получена за данните от Таблица 1 по МНК е показана на Фигура 3.

Нека си припомним, че регресионната линия се използва за предвиждане или оценяване на стойността на ***y*** по дадена стойност за ***x***. За всички стойности на ***x***, предвидените стойности на ***y***, означавани с $\hat{y}$, са разположени върху регресионната линия, която е показано на Фигура 3. Следователно, друг подход да се илюстрира метода на най-малките квадрати за получаване на уравнението на регресионната линия е като се разгледат грешките в предвиждането на ***y*** по ***x***. С други думи, чрез разглеждането на разликата между действителната стойност ***y*** и предвидената или оценена стойност $\hat{y}$. Очевидно, грешката на предвиждането ($y - \hat{y}$) е в действителност разстоянието между дадената точка и регресионната линия. Тогава, методът на най-малките квадрати е метод за минимизацията на сумата от квадратите $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Фигура 3. Регресионна линия на ***y*** върху ***x*** по МНК



По-нататък ще определим регресионния коефициент и свободния член на регресионното уравнение за прогнозиране на ***y*** по ***x.*** Ще ги означим $b_{y.x}$ и $a_{y.x}$, съответно. Специфичният процес на оценяването на стойности на $b_{y.x}$ и $a_{y.x}$, който използва МНК предполага използването на диференциално смятане и математически апарат, който е извън обсега на тази книга. Поради тази причина ще приведем съответните формули наготово, още повече, че поради големия обем изчисления използването на статистически пакет е почти задължително. Изразите за $b_{y.x}$ и $a_{y.x}$ са следните:

Регресионната линия (уравнение) се получава чрез изравняване на измерванията по метода на най-малките квадрати, който предполага да се намери минимума (най-малката стойност) на израза $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

(2) $$b_{y.x} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

(3) $$a_{y.x} = \frac{\sum_{i=1}^{n} y_i - b_{y.x}\sum_{i=1}^{n} x_i}{n} = \bar{y} - b_{y.x}\bar{x}$$

За илюстрация на използването на тези формули да разгледаме данните от Таблица 1:

$$b_{y.x} = \frac{20(3205) - (233)(257)}{20(3037) - (233)^2} = 0.65$$

$$a_{y.x} = \frac{257 - (0.65)(233)}{20} = 5.28$$

Следователно, регресионното уравнение за предвиждането на дебелината по ширината, оценено по 20 двойки измервания е:

$$\hat{y} = 0.65x + 5.28$$

Това уравнение може да бъде използвано за предвиждане по следния начин. Нека за даден артефакт е известно, че има дебелина 12 мм. Тогава можем да определим:

$\hat{y} = (0.65)(12) + 5.28 = 13.08$ мм.

Обратна (втора) регресионна линия

Връзката между двете случайни променливи може да бъде погледната и от друг ъгъл. Досега обсъждахме регресионното уравнение за предвиждането на y по x и съответните оценки $b_{y.x}$ и $a_{y.x}$. В по-голямата част от случайте това е естественият ред на предвиждане. Вече беше споменато, че стохастичните връзки не са еднозначно обратими. Това означава, че по регресионното уравнение, построено за предвиждане на y по x не е възможно да се реши обратната задача – да предвидим съответната стойност за **x** по зададена стойност за y чрез алгебраично обръщане на уравнението, както би могло да се направи при функционалната връзка. Следователно, нужно е да се построи второ регресионно уравнение, което да даде възможност да предвиждаме стойностите на променливата x по стойностите на променливата y.

За да предвидим стойностите на променливата y по стойностите на променливата x, регресионното уравнение е: $\hat{y} = b_{y.x}x + a_{y.x}$, ***където:*** $b_{y.x}$ ***се нарича регресионен коефициент, а*** $a_{y.x}$ ***се нарича свободен член или регресионна константа. Често това уравнение се нарича "проста линейна регресия".***

Тази регресионна линия може да бъде отново получена по метода на най-малките квадрати, но в случая ще бъде минимизиран изразът: $\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$. Регресионното уравнение за предвиждане на стойностите на променливата x по стойностите на променливата y ще бъде:

(4) $\hat{x} = b_{x.y}y + a_{x.y}$,

където:

(5) $$b_{x.y} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

(6) $$a_{x.y} = \frac{\sum_{i=1}^{n} x_i - b_{x.y}\sum_{i=1}^{n} y_i}{n} = \bar{x} - b_{x.y}\bar{y}$$

Ако използваме тези формули за данните от Таблица 1 ще получим:

$$b_{x.y} = \frac{20(3205) - (233)(257)}{20(3557) - (257)^2} = 0.83$$

$$a_{x.y} = \frac{233 - (0.83)(257)}{20} = 0.98$$

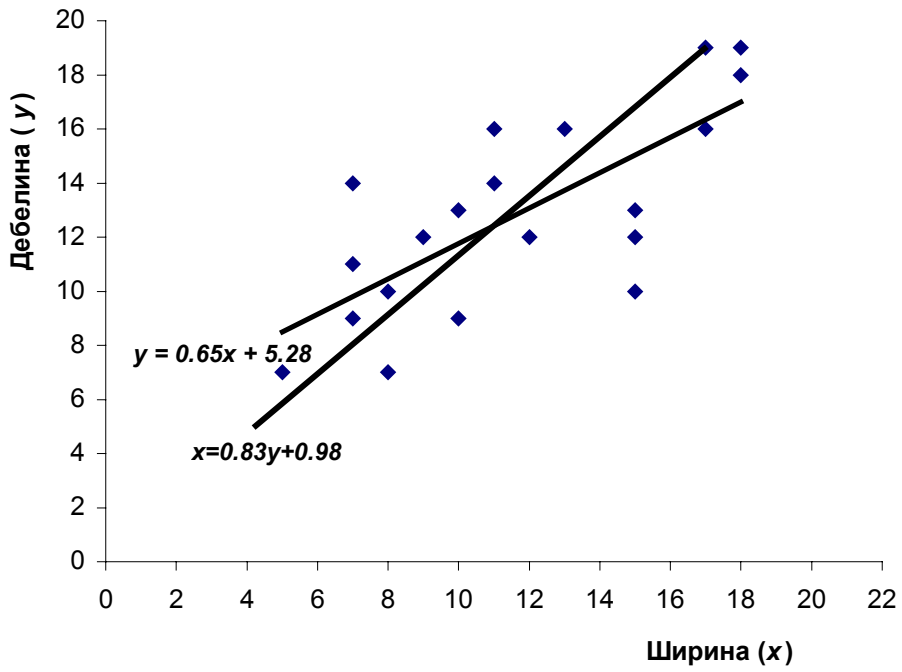
Следователно, регресионната линия за предвиждане на ширината по дебелината ще бъде $\hat{x} = 0.83y + 0.98$.

Ще обърнем внимание, че индексите на коефициентите $b_{x.y}$ и $a_{x.y}$ показват, че става дума за предвиждане на x по y, а не обратното, а също така показват, че за множеството от свързани по двойки данни могат да се построят две регресионни линии[2] в зависимост от това, коя променлива приемем за причина и коя – за следствие. Ще използваме полученото регресионно уравнение, за да определим стойността на x при $y = 12$: $\hat{x} = 0.83(12) + 0.98 = 10.94$.

За всяко множество от подвойкови измервания съществуват две различни регресионни линиии: една за предвиждане на y по x и една за предвиждане на x по y. В изследванията в поведенческите науки и науките за човека общопретият подход предполага да се предвижда стойността на y чрез извесна стойност на x.

На Фигура 4 са показани графично двете възможни регресионни линии за данните от Таблица 1.

Фигура 4. Графика на двете регегресионни линии



Предвидени стойности и тяхното разпределение

Един резултат от оценяването на уравнението на регресионната линия е, че сме в състояние да предвиждаме стойностите на y чрез стойностите на x. Типичен пример е предвижданото, да речем, на средния успех от следването по наблюдавани резултати от теста за интелегинтност IQ. Получаването на регресионното уравнение се основава на предварителни изследвания на взаимовръзката между тези две случайни променливи, за които е известно (не от статистиката, а от предметната област, от която идват данните), че се намират в причинно-следствени отношения. Ще отбележим, че

[2] Когато връзката между x и y е идеална, т.е. от стохастична преминава във функционолна, то всички точки ще лежат върху една права и, следователно, ще имаме само една регресионна линия.

след като веднъж е получено по представителна извадка, регресионното уравнение може да бъде прилагано върху всяка друга група от същата популация. Това е и смисълът на използването на регресионния анализ: да даде инструмент, с помощта на който да можем да предвидим очакваните стойности на зависимата променлива, ако са известни или достъпни само съответните стойности на независимата променлива. Очевидно, при този процес се допуска грешка (случайна), с която ще се занимаем по-подробно в следващия раздел. За да илюстрираме и обсъдим грешката при предвиждането ще разглеждаме както измерените (действителните) стойности на y, така и предвидените по модела стойности.

В примера за дебелината (y) и ширината (x) ни интересува предвиждането на резултата за дебелината по известен резултат за ширината, т.е регресията на y по x. Таблица 2 съдържа предвидените стойности за 20 артефакта, както и друга информация, която ще ни е необходима по-нататък. Графично, тези предвидени стойности лежат на една линия, изобразена на Фигура 3. Интересно е да отбележим, че предвидената стойност на y за средната стойност на x е средната на y или: $\bar{y} = b_{y.x}\bar{x} + a_{y.x}$. Това се потвърждава и от данните в Таблица 1:

$$\mathbf{12.85 = (0.65)(11.65) + 5.28}$$

Таблица 2. Предвидени резултати и грешки при предвиждането на резултата за дебелината (y) на базата на резултата за ширината (x)

Артефакт	x	y	$\hat{y}$	$\varepsilon = y - \hat{y}$	$\varepsilon^2 = (y - \hat{y})^2$
1	15	12	15.03	-3.03	9.18
2	10	13	11.78	1.22	1.49
3	7	9	9.83	-0.83	0.69
4	18	18	16.98	1.02	1.04
5	5	7	8.53	-1.53	2.34
6	10	9	11.78	-2.78	7.73
7	7	14	9.83	4.17	17.39
8	17	16	16.33	-0.33	0.11
9	15	10	15.03	-5.03	25.30
10	9	12	11.13	0.87	0.76
11	8	7	10.48	-3.48	12.11
12	15	13	15.03	-2.03	4.12
13	11	14	12.43	1.57	2.46
14	17	19	16.33	2.67	7.13
15	8	10	10.48	-0.48	0.23
16	11	16	12.43	3.57	12.74
17	12	12	13.08	-1.08	1.17
18	13	16	13.73	2.27	5.15
19	18	19	16.98	2.02	4.08
20	7	11	9.83	1.17	1.37
Общо				**0.00**	**116.59**

Да обсъдим разпределението на предвидените резултати и тяхната средна стойност $\bar{\hat{y}}$. Всяка предвидена стойност се получава чрез равенството $\hat{y}_i = b_{y.x}x_i + a_{y.x}$, i**=1, 2 ..,**n. Сумирайки за всички измервания и делейки резултата на n ще имаме:

$$\frac{\sum_{i=1}^{n}\hat{y}_i}{n} = \frac{\sum_{i=1}^{n}(b_{y.x}x_i + a_{y.x})}{n}.$$

Лявата страна на това равенство по определение е средната стойност на разпределението на предвидените стойности. Чрез несложни преобразования получаваме:

Разпределението на предвидените стойности има средна, равна на средната на действително измерените стойности, т.е. $\bar{\hat{y}} = \bar{y}$. Графично, всички предвидени стойности лежат върху регресионната линия.

$$\bar{\hat{y}} = b_{y.x}\frac{\sum_{i=1}^{n}x_i}{n} + \frac{\sum_{i=1}^{n}a_{y.x}}{n},$$

но $\sum_{i=1}^{n}x_i / n = \bar{x}$, а $\sum_{i=1}^{n}a_{y.x} / n = a_{y.x}$. Следователно, $\bar{\hat{y}} = b_{y.x}\bar{x} + a_{y.x}$.

Тъй като по-горе вече получихме, че $\bar{y} = b_{y.x}\bar{x} + a_{y.x}$, то $\bar{\hat{y}} = \bar{y}$. Този резултат ще бъде използван по-късно, когато обсъждаме подробно грешката на предвиждането. За емпиричната проверка на този резултат можем да използваме данните от Таблица 1. Ще получим: $\bar{\hat{y}} = 12.85 = \bar{y}$.

Грешки на предвиждането

След като знаем разпределението на предвидените стойности можем да сравняваме една действителна индивидуална стойност с предвидената стойност. Разликата между тези две стойности се определя като *грешка на предвиждането* (много често се използва и термина *грешка на оценката*) и се бележи като $\varepsilon_i = (y_i - \hat{y}_i)$. Грешките на предвиждането на резултатите за дебелината по резултатите от ширината са показани в Таблица 2. Неправдоподобно е измерената и предвидената индивидуални стойности да съвпаднат, но може да се случи. В този случай грешката на предвиждането ще бъде нула. Когато разглеждаме предвидените стойности и грешката на предвиждането, не можем да очакваме точното предвиждане на всеки отделен резултат или на група резултати. По-скоро нашите очаквания са свързани с цялостното разпределение на грешките и, по-специално, с дисперсията и стандартното отклонение на това разпределение. Затова първо ще отговорим на въпроса каква е средната стойност на разпределението на грешките. Да запишем:

Грешката в предвиждането е разликата между действителната и предвидената сотйности на индивидуалното измерване. Средната на разпределението на грешките е равна на нула.

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

След несложни алгебраични преобразования ще получим:

$$\bar{\varepsilon} = \frac{1}{n}\sum_{ui=1}^{n}y_i - \frac{1}{n}\sum\hat{y}_i = \bar{y} - \bar{\hat{y}}$$

Тъй като вече установихме, че $\bar{y} = \bar{\hat{y}}$, то $\bar{\varepsilon} = 0$. Това свойство може да бъде лесно проверено чрез изчисляването на средната на грешките от Таблица 2.

Стандартна грешка на предвидената стойност (оценката)

Сега ще разгледаме начина на изчисляване на дисперсията и стандартната грешка на разпределението на предвидените стойности. Стандартното отклонение на това разпределение често се нарича *стандартна грешка на оценката*. Формално, извадковата дисперсия на разпределението[3] на грешките може да се запише:

Стандарнтоно отклонение на разпределинето на грешките на оценяването се нарича стандарнтна грешка на оценката и се изчислява в съответствие с формула (8).

(7) $$s^2_{y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2$$

Стандартната грешка на оценката $s_{y.x}$ е $\sqrt{s^2_{y.x}}$. Тъй като показахме, че средната на грешките е нула, то формулата може да се препише във вида:

(8) $$s_{y.x} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2}$$

Ако използваме данните от Таблица 2, то ще получим: $s^2_{y.x} = 116.59/18 = 6.48$ и $s_{y.x} = \sqrt{6.48} = 2.55$.

Допускания при построяването на регресионната линия

Построяването и използването на регресионната линия предполага, че са изпълненне някои предварителни предположения. Тези предположения се обясняват по-долу и графично се илюстрират на Фигура 5. Изискванията, които трябва да са налице относно x и y в популацията (когато предвиждаме y чрез x) са следните:

Съществуват няколко условия, или предположения, свързани с регресионната линия. Едно от най-важните е наличието на хомоскедестичност. Хомоскедастичността означава равенство на дисперсиите на грешките от предвиждането за всяка възможна стойност на x от популацията.

1. x и y са непрекъснати променливи.
2. x и y се измерват поне в интервална скала
3. Връзката между x и y е *линейна*.

Предположенията относно грешките на предвиждането ε_i са следните:

1. Грешките са нормално разпределени със средна нула и еднакви дисперсии, равни на $\sigma^2_{y.x}$.
2. Дисперсията на грешките на предвиждането е една и съща за всяка стойност на x. Това изискване обикновено се означава като *хомоскедастичност*. За всяка стойност на x предвидената стойност за y лежи на регресионната линия.

[3] В глава 7 определихме степените на свобода като броя на измерванията минус броя на ограниченията, наложени върху тях. При определянето на стандартната грешка на оценката, налице са две ограничения и, следователно, губим две степени на свобода. Едната е свързана с оценяването на регресионния коефициент $b_{y.x}$, а другата е свързана с оценяването на константата $a_{y.x}$. Следователно, знаменателят в израза за оценявато на $\sigma^2_{y.x}$ трябва да е $n-2$.

Ако предположим, че са налице голям брой подвойкови измервания (x_i, y_i; i=1,2,…,n) и нека всички обекти имат равни x_i, но различни y_i. В този случай ще се наблюдава картината, представена графично на Фигура 5. Налице е разпределение на стойностите на грешките около оценената стойност $\hat{y}$. Разпределението на грешките за тази частна стойност на x (също както разпределението на грешките за всяка стойност на x) се предполага, че е нормално. В допълнение към предположението за нормалност, ако се изиска и хомоскедастичнот, то дисперсиите на всички грешки са равни по между си и са равни на $\sigma^2_{y.x}$.

Фигура 5. Графична илюстрация на допусканията в модела на линейна регресия

Връзка между корелацията и регресията

+Дотук разглеждахме регресионното уравнение и развихме концепцията за стандартната грешка на оценката. Сега ще се обърнем към въпроса за връзката между *корелацията* и *предвиждането*. Да разгледаме за дадена стойност на x действително измерената сотйност на y. В термините на предишните обсъждания (например, при дисперсионния анализ) това измерване се представя като съставено от две части. Първата част е оценената стойност $\hat{y}$, втората част е грешката на предвиждането $(y - \hat{y})$. Формално този подход се изразява чрез равенството:

(9) $$y = \hat{y} + (y - \hat{y})$$

Разбиване на дисперсията на зависимата променлива

Да разгледаме общата дисперсия на измерванията на зависимата променлива y, означавана като s_y^2. Разликата между измерената стойност и средната за y може да се разбие на две части:

(10) $$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}), \; i = 1,2,...,n$$

Този начин на представяне е илюстриран на Фигура 6. Повдигайки на квадрат двете страни на равенството и сумирайки за всички измерени стойности ще намерим:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}[(\hat{y}_i - \bar{y}) + (y_i - \hat{y})]^2$$

$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Тъй като $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$, то вторият член в дясната страна отпада и ще имаме:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Всеки от изразите в това равенство се явява числител във вече определените по-рано компоненти на дисперсията. Като разделим всеки от тези изрази на свързаните с него степени на свобода ***(n-2)*** ще получим, че общата дисперсия може да се представи като сума от две събираеми. Формално:

Общата дисперсия на измерванията на случайната променлива y може да се разбие на две части: дисперсия на предвидените стойности и дисперсия на грешките на оценките.

(11) $$s_y^2 = s_{\hat{y}}^2 + s_{y.x}^2 \; ,$$

където: s_y^2 = общата дисперсия на y

$s_{\hat{y}}^2$ = дисперсията на предвидените стойности $\hat{y}$

$s_{y.x}^2$ = дисперсията на грешкине на оценката.

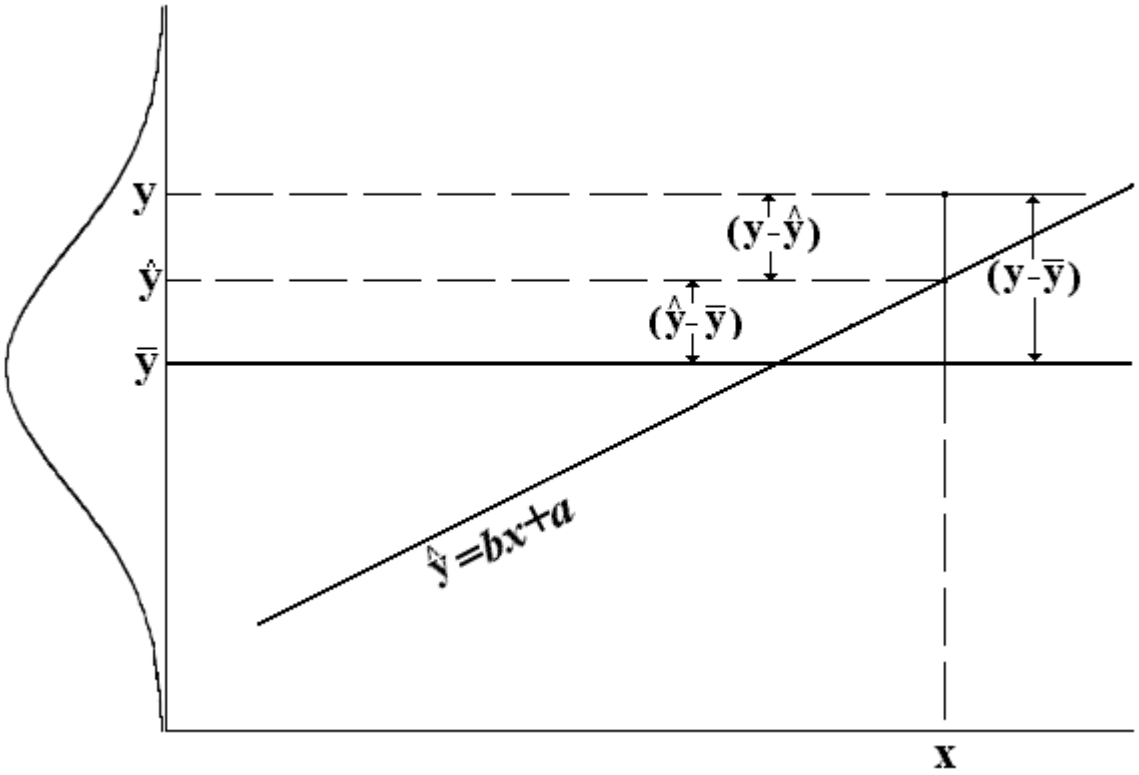
Нека разделим двете страни на равенство (11) на $s_y^2 (\neq 0)$:

(12) $$\frac{s_y^2}{s_y^2} = 1 = \frac{s_{\hat{y}}^2}{s_y^2} + \frac{s_{y.x}^2}{s_y^2}$$

Да разгледаме израза $s_{\hat{y}}^2 / s_y^2$. Това е отношението на дисперсията на предвидените стойности и общата дисперсия. С други думи, това е частта от общата дисперсия на y, която може да бъде обяснена посредством знанието на съответната стойност на x (чрез предвиждането на y чрез x). В глава 5 установихме, че по този начин се определя *коефициентът на детерминация* r^2. Следователно горният израз става:

$$1 = r^2 + \frac{s_{y.x}^2}{s_y^2}.$$

Фигура 6. Разбиване на $(y_i - \bar{y})$ на две съставящи: $(\hat{y}_i - \bar{y})$ и $(y_i - \hat{y}_i)$



По-нататъшните преобразования на този израз водят до една опростена формулата за изчисляване на дисперсията на грешката на оценката $s^2_{y.x}$ и на стандартната грешка на оценката $s_{y.x}$:

(13) $$s^2_{y.x} = s^2_y(1 - r^2)$$

(14) $$s_{y.x} = s_y\sqrt{1 - r^2}$$

Формула (14) показва, че ако корелацията между x и y е идеална (+1.0 или –1.0), то не съществува дисперсия, която да се дължи на грешката на оценката. Общата дисперсия ще бъде равна на обяснената дисперсия и всички измерени и предвидени стойности ще съвпадат. Стандартната грешка на оценката ще е нула. От друга страна, ако корелацията е нула, то необяснената дисперсия, или дисперсията на грешката, ще е равна на наблюдаваната дисперсия. Стандартната грешка на оценката ще е равна на стандарното отклонение на разпределението на измерванията на y. Следователно, когато корелацията е нула, независимата променлива не може да обясни никаква част от дисперсията на зависимата променлива. Обяснението тук не предполага обезателно наличието на причинност. То само показва как се разпределя общата дисперсия. В глава 5 корелационния коефициент беше определен като индекс, който показва силата на *функционална* връзка между две променливи. Дали тази връзка може да разглеждана като причинна е предмет на интерпретацията на естеството на променливите, които корелират.

Отношението на дисперсията на предвидените измервания към общата дисперсия на y е равно на r^2, коефициента на детерминация. Казано по друг начин, r^2 е равен на частта от дисперсията на променливата y, която се обяснява чрез информацията, която имаме за променливата x.

Корелация и регресионни коефициенти

Нека първо отбележим, че между формулите за изчисляване на регресионните коефициенти $b_{y.x}$ и $b_{x.y}$, от една страна и формулата за изчисляване на корелационния коефициент, от друга, съществува определена прилика. За да разкрием връзката между тези два типа коефициенти, най-напред ще трябва да въведим една мярка на зависимост между две променливи, наречена *ковариация*. Ковариацията се различава от корелационния коефициент по това ,че може да приема стойности и извън интервала [-1, +1]. Ковариацията за една крайна популация, означавана често като σ_{xy} се определя като:

Ковариацията е мярка за връзката между две променливи и се определя като средната на произведенията на отклоненията на измерванията на двете променливи от техните средни аритметични. В термините на това определение връзката между корелационния коефициент и двата възможни регресионни коефициенти може лесно да бъде показана.

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N} = \frac{\sum_{i=1}^{N} x_i' y_i'}{N}, \tag{15}$$

където: $x_i' = x_i - \mu_x$, $y_i' = y_i - \mu_y$

Аналогично, неизместената извадкова оценка (в съответсвие с обсъждането в глава 3) за ковариацията, означавана като s_{xy}, ще се определя като:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i' y_i'}{n-1}. \tag{16}$$

Както може да се види, ковариацията е средна от произведението на отклоненията на x и y от техните средни. Съгласно тези определения, формулите за корелационния коефициент и за двата регресионни коефициента могат да бъдат представени в термините на ковариацията:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\textit{ковариация}}{\textit{произведение на стандартните отклонения}}$$

$$b_{y.x} = \frac{s_{xy}}{s_x^2} = \frac{\textit{ковариация}}{\textit{дисперсия на } x}$$

$$b_{x.y} = \frac{s_{xy}}{s_y^2} = \frac{\textit{ковариация}}{\textit{дисперсия на } y}$$

Връзката между между корелационния коефициент и регресионните коефициенти се дава от изразите:

$$b_{y.x} = r\frac{s_y}{s_x} \Rightarrow r = b_{y.x}\frac{s_x}{s_y} \tag{17}$$

$$b_{x.y} = r\frac{s_x}{s_y} \Rightarrow r = b_{x.y}\frac{s_y}{s_x}. \tag{18}$$

Като умножим горните две равенства ще получим:

$$r^2 = b_{y.x} b_{x.y}. \tag{19}$$

Следователно, квадратът на корелационния коефициент е равен на произведението от двата регресионни коефициента.

Предвиждане на стандартните стойности на у по стандартните стойности на х

В глава 4 стандартната стойност за извадковите измервания беше оперделена чрез израза:

$$z_i = \frac{x_i - \bar{x}}{s_x}, \; i = 1,2,...,n$$

Да предположим, че е необходимо да построим регресионно уравнение за предвиждане на стандартните стойности на y, означавани като z_y, по съответните стандартни стойности за x, означавани като z_x. Да разгледаме регресионното уравнение за y_i: $\hat{y}_i = b_{y.x} x_i + a_{y.x}$, където: $a_{y.x} = \bar{y} - b_{y.x}\bar{x}$. Като заместим константата в горното уравнение с този израз ще получим:

$$\hat{y}_i = \bar{y} + b_{y.x}(x_i - \bar{x}).$$

Тъй като $b_{y.x} = r(s_y / s_x)$, то

$$\hat{y}_i = \bar{y} + r\frac{s_y}{s_x}(x_i - \bar{x}).$$

След несложни алгебраични преобразования получаваме:

(20) $$\frac{\hat{y}_i - \bar{y}}{s_y} = r\left(\frac{x_i - \bar{x}}{s_x}\right) \Rightarrow z_{\hat{y}} = rz_x.$$

Може по същия начин да се покаже, че за предвиждането на стандартните стойности z_x по съответните стандартни стойности z_y, регресионната уравнение ще има вида:

(21) $$z_{\hat{x}} = rz_y.$$

Следователно, когато се предвижда стандартна стойност чрез друга стандартна стойност, регресионният коефициент (наклонът на регресионната линия) е един и същ при предвиждането на y по x и при предвиждането на x по y. Регресионният коефициент е равен на корелационният коефициент. Този резултат лесно следва и от факта, че стандартните стойности имат средна на разпределинието нула и дисперсия единица.

Вероятности, свързани с регресията и предвиждането

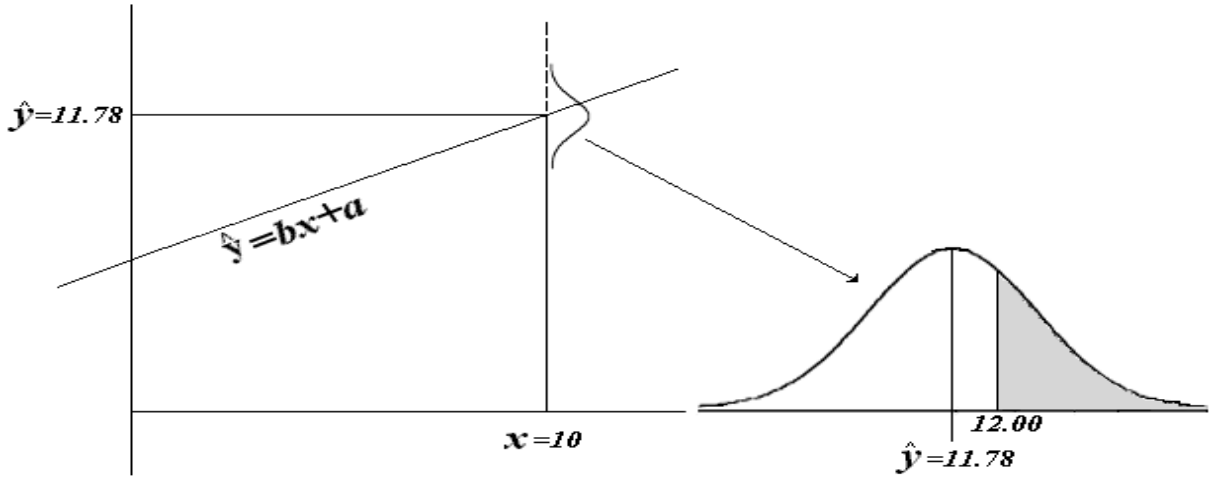
В глава 4 бяха обсъдени стандартните стойности и в тази връзка, лицето на повърхнината под кривата на плътностната функция на стандартното нормално разпределение. Тук няма да се спираме на въпроса каква е вероятността една наблюдавана и съответната оценена стойности да съвпаднат. Очевидно, при непрекъснатите разпределения, каквото е нормалното, тази вероятоност е нула. Поради това, ще се интересуваме от вероятността при дадена стойност на независимата променлива каква е вероятността истинската стойност на зависимата стойност да лежи в даден интервал. За да развием този въпрос, ще се върнем към примера по-горе.

Да предположим, че i-тия артефакт има ширина (x) 10 мм. Каква е вероятността, че този артефакт да има дебелина 12 и повече мм? Първо ще получим предвидената стойност или оценката:

$$\hat{y}_i = 0.65x_i + 5.28 = 0.65(10) + 5.28 = 11.78$$

Следователно, за ширина от 10 мм, най-добрият предвиден (прогнозен) резултат за дебелината е 11.78 мм. Нека си припомним, че от допусканията на регресионния анализ следва, че за всяко фиксирано измерване на променливата x, съществува разпределение на наблюдаваните стойности на y около предвидената стойност (оценката) $\hat{y}$. В този пример за x = 10, разпределението на измерванията на y има средна, равна на 11.78, т.е. $\hat{y}$, и стандартно отклонение 2.55 (стандартната грешка на оценката). На Фигура 7 е показано разпределението на y за x = 10 и стандартната грешка на оценката.

Фигура 7. Илюстрация на вероятността, свързана с предвиждането



Вероятността, че съответното измерване на y ще е равно или по-голямо от 12 при това разпределение съответства на относителната честота на измерванията от това разпределение, които са равни и по-големи от 12. С други думи, вероятността е равна на лицето на повърхнината под нормалната крива (със средна = 11.78 и стандартно отклонение = 2.55), което лежи в дясно от стойността 12. Това лице (защрихованата част на разпределението на Фигура 7) може да се намери по следния начин, който вече приложихме в подобна ситуация, превръщайки съответната стойност към стандартна и използвайки таблицата за стандартното нормално разпределение. Общата формула беше:

Вероятност на предвиждането, това е вероятността, че предвидената (оценената) стойност ще попадне в определен интервал при фиксирана стойност на независимата променлива, дадено регресионно уравнение и стандартна грешка на оценката.

$$z_i = \frac{\textit{измерване} - \textit{средна}}{\textit{стандартно отклонение}}.$$

В нашия случай:

$$\frac{y_i - \hat{y}}{s_{y.x}} = \frac{12 - 11.78}{2.55} = 0.086$$

Таблица B1 показва стойностите на лицата, определени за точките **вляво** от дадената, т.е. $p(z \le z_i)$ Затова нужното ни лице ще се получи като извадим табличната стойност от 1 или $p(z \ge z_i) = 1 - p(z \le z_i)$. В примера имаме: $p(y \ge 12) = p(z \ge 0.086) = 1 - p(z \le 0.086)$. От Таблица B1 имаме: за 0.08 лицето е 0.5319, а за 0.09 е 0.5359. Можем да вземем средното лице, или 0.5339. Тогава, нужната вероятност ще е 1-0.5359 = 0.4641. Следователно, вероятността, че дебелината ще е по-голяма от 12 мм, когато резултата ширината е 10 мм, е 0.4641. Съответно, можем по подобен начин да определим вероятността дебелината да лежи между всеки две точки (или под и над всяка точка) върху скалата на измерване на резултата.

Доверителни интервали на предвиждането

Да разгледаме отново примера за предвиждането на дебелината, ако ширината е 10 мм. Да предположим, че трябва да построим такъв интервал около оценката, че да има вероятност 0.95 в него да попадне истинската стойност. Този интервал нарекохме *95 процентен доверителен интервал* и го означихме 95%ДИ. Процедурата включва използването на разпределението на действителните измервания на y за дадена стойност на x, както е показано на Фигура 7. Предположението е, че наблюдаваните измервания на y, свързани с определена стойност на x, са нормално резпределении със средна, равна на $\hat{y}$ и стандартно отклонение, равно на $s_{y.x}$. Тъй като 95% от цялото лице на повърхнината под стандартната нормална крива се намира между –1.96 и +1.96 стандартни отклонения около средната, то 95%ДИ ще има вида:

$$95\%ДИ = 11.78 \pm 1.96(2.55) = 11.78 \pm 5.00 = (6.78, 16.78)$$

Вече можем да кажем с увереност 95 процента, че ако ширината на артефактът е 10 мм, то индивидуалната дебелина ще се намира между точките 6.78 и 16.78.

95%ДИ е интервалът, в който се очаква с увереност 95% да се намира истинската, но неизвестна стойност на променливата y при задена стойност на променливата x.

Възможно е да се определят доверителни интервали с вероятност, различна от 0.95. Основанията да използваме 95%ДИ, също както и за всеки друг интервал, бяха разгледани в някои от предходните глави. Възможно е също да се построят доверителни интервали за предвидените стойности $\hat{y}$ за различни стойности на независимата променлива x. Ако условието за хомоскедастичност е изпълнено, то общата формула за 95%ДИ е:

(22) $$95\%ДИ = \hat{y} \pm 1.96 s_{y.x}$$

Резюме

Предвиждането е важно практическо приложение на концепцията за корелация. В тази глава беше описан процеса на предвиждане на една променлива чрез знанието на стойностите на друга променлива посредством линейна регресия. Едно от необходимите условия е променливите да са свързани линейно.

Уравнението, чрез което се получава предвидената стойност е регресионно равенство, което свързва двете променливи x и **y**. В действителност има две регресионни линии, които свързват горните променливи. Това се вижда добре от диаграмата на разсейване.В повечето ситуации по-полезна е правата регресия, т.е. регресията на **y** върху x. Регресионното уравнение се получава по метода на най-

малките квадрати, който минимизира дисперсията на грешките между наблюдаваните и предвидените стойности. Уравнението, по което се извършва предвижданото е уравнение на права линия, зададено с формулата $\hat{y} = bx + a$.

С нарастването на абсолютната стойност на корелационния коефициент между зависимата и независимата променливи грешката на оценката намалява. Възможността да предвижда и възможността да се свърже резултата от това предвиждане с определена вероятност не трябва да се смесва с причинно-следственото обяснение на връзката между двете променливи. Резултатите от общия тест са тясно свързани с резултатите от теста по математика. Таво не означава, че ***x*** е причина за **y,** или обратно, че **y** е причина за ***x***. Статистиката единствено потвърждава, че има прогнозна връзка между двете променливи.

Корелационният коефициент дава количествен израз на връзката между двете променливи. Целта на регресията е *предвиждането*. Възможността да се предвижда с определена степен на точност не е маловажна, а служи за постигането на определени цели. Когато изследователят е постигнал дадена цел чрез регресионния анализ, едно разбиране как променливите си взаимодействат в причинно-следствен план може вече да бъде получено на базата на специфичния контекст на задачата.