Глава 16

Анализ на зависимости

16.1 Видове зависимости в статистиката

В много задачи е необходимо да се установи и оцени зависимостта на изучавана случайната величина Y (признак) от една или няколко други случайни величини $X_1, X_2, \ldots, X_s$ (фактори).

Зависимостта между случайните величини в реални условия може да бъде различна. В някои случаи случайните величини X и Y може да са **независими** (вж. разд. 8.4, 15.3). В други случаи зависимостта между X и Y може да е толкова силна, че ако знаем каква стойност е приела едната величина X, може точно да укажем стойността на другата Y. Придържайки се към традиционната терминология, може да кажем, че зависимостта между X и Y е **функционална**. С примери на такива зависимости често се срещаме в природата и техниката.

В същото време може да се посочат и примери от друг вид – когато зависимост между случайните величини съществува, но няма строго изразен функционален характер. Подобни примери са особено характерни за такива области на науката и практиката като биология, медицина, агротехника, икономика и др., където развитието на различни процеси и явления, като правило, зависи от много фактори, които трудно може да се отчетат в своята пълнота.

Известно е например, че обилните валежи в периода на съзряване на пшеницата води до повишаването на добива. Обаче това не означава, че връзката между количеството валежи X (литри/кв.м.) и добива на пшеница Y (кг/дка) е функционална; Освен валежите на добива оказват

влияние и други фактори: тип на почвата, количествата на внесените торове, броят на слънчевите дни през периода на вегетацията и др. В подобни ситуации, когато изменението на една величина влияе на друга само статистически (усреднено), е прието да се говори за **статистическа** зависимост между величините. В частност, статистическата зависимост се проявява в това, че при изменението на едната от величините се изменя средната стойност на другата; в този случай статистическата зависимост се нарича **корелационна**.

Ще дадем няколко примера, които илюстрират разни степени на зависимост между случайни величини.

Пример 16.1. *Нека U е електрическото напрежение, приложено в краищата на електрическа верига, а I е силата на тока, протекъл през веригата при това напрежение. Оказва се, че зависимостта между тези величини е функционална и се изразява със закона на Ом: $U = RI$, където R е съпротивлението на веригата, което не зависи от I и U.*

Пример 16.2. *Нека X (метри) е височината на случайно избрано дърво в борова гора, а Y (сантиметри) е неговият диаметър при основата. Тук зависимостта не е функционална, но може да се счита за силна. Ако наблюденията на тези две величини се извършват в смесена гора (бор, бреза, трепетлика), то ще имаме средна корелационна зависимост.*

Пример 16.3. *Нека от купчина камъни с неправилна форма се избира случайно един камък и се измерват теглото му X (кг) и най-голямата му дължина Y (см). Зависимостта между X и Y е слабо изразена.*

Пример 16.4. *Нека X (см) е ръстът на случайно избран възрастен мъж, а Y (години) е неговата възраст. Наблюденията показват, че тези величини практически са независими.*

16.2 Методи за анализ

Изследването на взаимовръзката между факторите $X_1, X_2, \ldots, X_s$ и признака Y се извършва със статистическите методи: корелационен анализ и регресионен анализ.

Задачата на **корелационния анализ** е да установи степента на влияние на факторите $X_1, X_2, \ldots, X_s$ върху признака Y. Корелационният

анализ позволява да се проявят неизвестните връзки между факторите и признака, да се определят главните компоненти – факторите, които оказват най-голямо влияние върху изменението на стойностите на признака. В някои случаи, в резултат на корелационния анализ може да се установи типа на зависимост между факторите и признака: линейна, степенна, експоненциална, логаритмична и др.

След извършването на корелационния анализ и при някои допълнителни предположения се избира подходящ **математически модел**, който включва т.нар. **уравнение на регресия**

$$y = f(x_1, x_2, \ldots, x_s; a_1, a_2, \ldots, a_k),$$

където f се нарича **функция на регресия**, а параметрите $a_1, a_2, \ldots, a_k$ са неизвестни и предстои да бъдат определени.

Ако функцията на регресия е линейна относно параметрите (но не задължително относно факторите), то се говори за **линеен модел на регресия**. В противен случай моделът на регресия се нарича **нелинеен**.

Чрез регресионния анализ се доуточнява избраната функция на регресия, като параметрите $a_1, a_2, \ldots, a_k$ се подберат така, че функцията на регресия да отразява възможно най-добре взаимната връзка между факторите $X_1, X_2, \ldots, X_s$ и признака Y, съдържаща се в статистическите данни за тези величини.

Статистическите проблеми на регресионния анализ са:

• получаване на най-добри точкови и интервални оценки за неизвестните параметри на регресията;

• проверка на хипотези относно тези параметри;

• проверка на адекватността на предложения модел;

• проверка на направените предположения.

Регресионният анализ се извършва по две причини. Първо, описанието на зависимостта между $X_1, X_2, \ldots, X_s$ и Y помага да се установи наличието на възможна причинна връзка между тях. Второ, уравнението на регресия позволява да се предсказват стойностите на Y по получените стойности на факторите. Тази възможност е особено важна в случаите, когато непосредствените измервания на Y са трудни за осъществяване или са скъпо струващи.

Следващите ни разглеждания са свързани с определянето на зависимостта на една случайна величина (признак) Y от един фактор X, като по-подробно ще се спрем на случая на **проста линейна регресия**.

16.3 Корелационен анализ

16.3.1 Корелационно поле

Наличието на зависимост между две случайни величини X и Y се установява въз основа на съвместна извадка от стойности на (X, Y):

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\,. \tag{16.1}$$

Нанасяме точките $(x_i, y_i), \quad i = 1, 2, \ldots, n$ в равнината Oxy. Полученото множество от точки се нарича **корелационно поле (диаграма на разсейването, SCATTER PLOT)**.

Корелационното поле е много показателно при действително наличие на функционална зависимост между величините, защото тогава точките $(x_i, y_i),\ i = 1, 2, \ldots, n$ са близко разположени около графиката на зависимостта. В противен случай точките са разположени хаотично.

На фиг. 16.1 са дадени примерни разположения на точките на корелационното поле при различни видове на зависимост между X и Y:

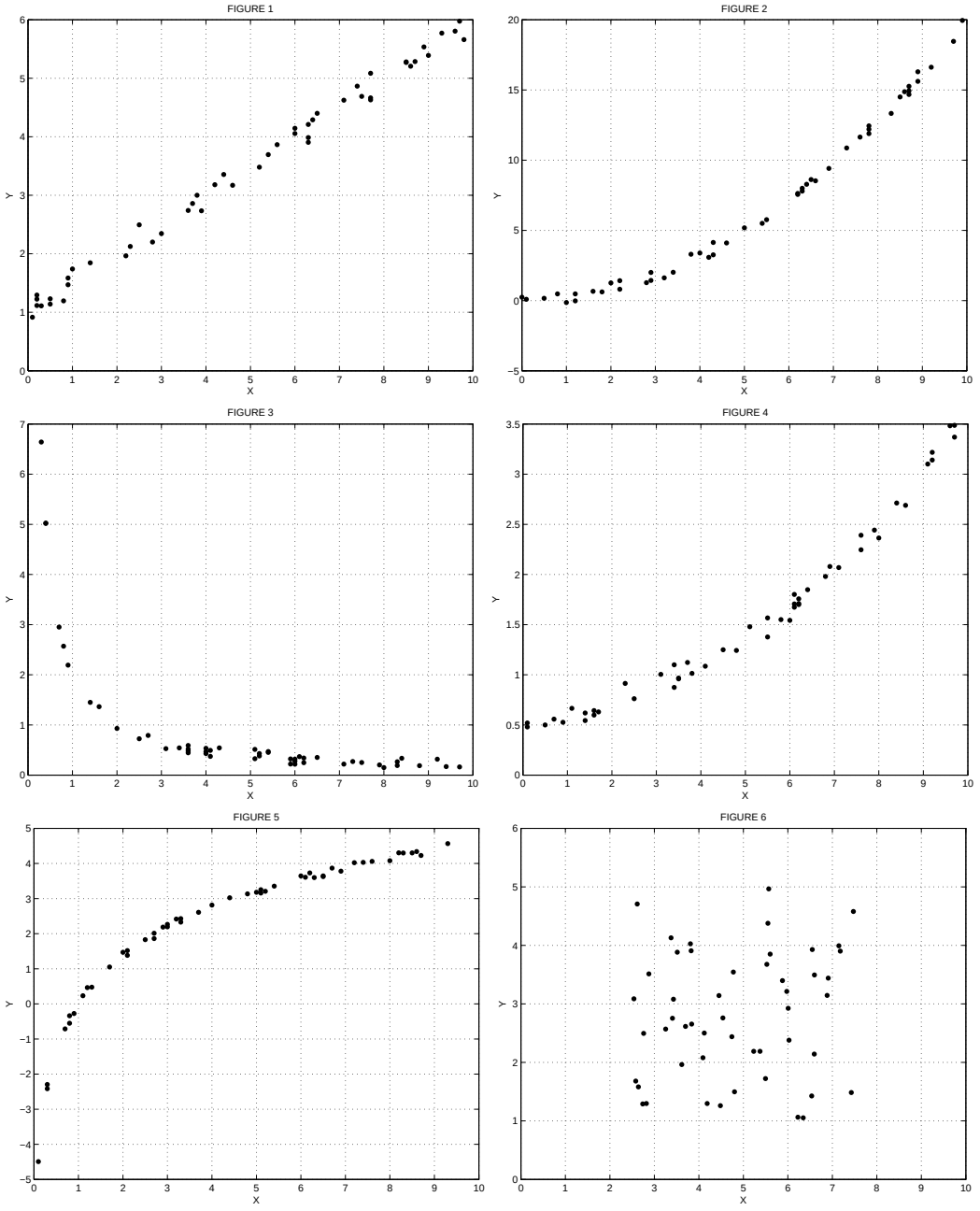
- линейна зависимост (FIGURE 1);
- степенна зависимост (FIGURE 2);
- обратна пропорционалност (FIGURE 3);
- експоненциална зависимост (FIGURE 4);
- логаритмична зависимост (FIGURE 5);
- липса на функционална зависимост (FIGURE 6).

16.3.2 Линейна корелация. Статистически коефициент на корелация

Най-простата и най-лесна за изучаване връзка между величините X и Y е линейната

$$Y = aX + b. \tag{16.2}$$

В случай на линейна връзка точките на корелационното поле се групират в по-голяма или по-малка степен около права линия.



Фиг. 16.1: Корелационни полета

Както видяхме в раздел 8.6, показател за наличието на линейна корелация (освен корелационното поле) е коефициентът на корелация r_{XY} в случая, когато $|r_{XY}| = 1$.

Ако $r_{XY} = +1$, то имаме линейна корелация и $a > 0$ (положителна корелация);

Ако $r_{XY} = -1$, то имаме линейна корелация и $a < 0$ (отрицателна корелация).

На практика не знаем r_{XY}, но можем да го оценим чрез извадката (16.1), като пресметнем **статистическия коефициент на корелация** $\overline{r}_{XY}$. За целта пресмятаме величините

$$S_x = \sum_{i=1}^{n} x_k, \quad S_y = \sum_{i=1}^{n} y_i \,, \tag{16.3}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2, \quad S_{yy} = \sum_{i=1}^{n} y_i^2, \quad S_{xy} = \sum_{i=1}^{n} x_i y_i \,, \tag{16.4}$$

след което определяме оценките за EX, EY, DX, DY, K_{XY} и r_{XY}:

$$\overline{x} = \frac{S_x}{n}, \quad \overline{y} = \frac{S_y}{n}, \tag{16.5}$$

$$\overline{D}_X = \overline{\sigma}_X^2 = \frac{nS_{xx} - S_x^2}{n^2}, \quad \overline{D}_Y = \overline{\sigma}_Y^2 = \frac{nS_{yy} - S_y^2}{n^2}, \tag{16.6}$$

$$\overline{K}_{XY} = \frac{nS_{xy} - S_x S_y}{n^2}, \tag{16.7}$$

$$\overline{r}_{XY} = \frac{\overline{K}_{XY}}{\overline{\sigma}_X \overline{\sigma}_Y} = \frac{\overline{K}_{XY}}{\sqrt{\overline{D}_X \overline{D}_Y}} \cdot \tag{16.8}$$

Счита се, че:

при $|\overline{r}_{XY}| < 0.3$ корелационната зависимост е слаба;

при $|\overline{r}_{XY}| = 0.3 \div 0.7$ корелационната зависимост е средна;

при $1 \geq |\overline{r}_{XY}| > 0.7$ корелационната зависимост е силна.

Ако $\overline{r}_{XY} > 0$, то имаме положителна корелация.

Ако $\overline{r}_{XY} < 0$, то имаме отрицателна корелация.

Статистическият коефициент на корелация $\overline{r}_{XY}$ може да приеме стойност 1 или -1 само, ако точките (x_i, y_i) лежат точно върху права линия. Това е възможно, когато между X и Y има детерминирана линейна връзка. Ще отбележим обаче, че дори когато връзката между величините X и Y е строго функционална ($Y = f(X)$), точките (x_i, y_i) не лежат точно

върху графиката на функцията $y = f(x)$, понеже стойностите x_i и y_i са получени в резултат на измерване, което винаги е свързано с допускане на грешки.

С подходящи смени на променливите някои функционални зависимости се свеждат до линейна зависимост. Например:

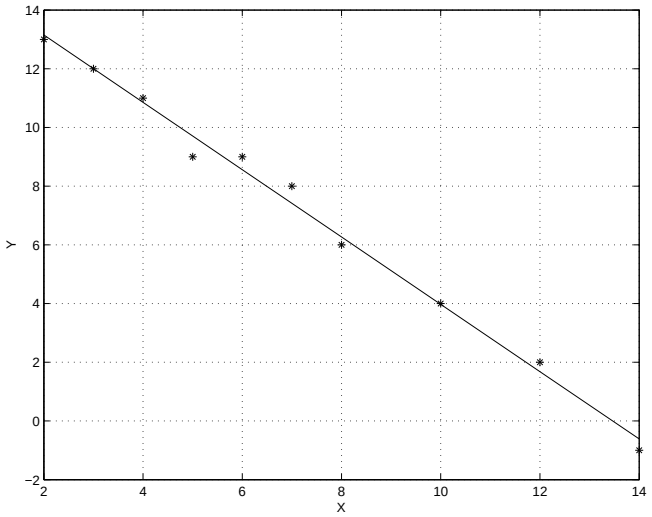
- $y = \dfrac{1}{ax+b}$. Полагаме $z = \dfrac{1}{y}$ и получаваме $z = ax + b$;
- $y = ce^{\mu x}$. Полагаме $z = \ln y$ и получаваме $z = \mu x + \ln c$;
- $y = a \ln x + b$. Полагаме $u = \ln x$ и получаваме $y = au + b$;
- $y = cx^a$. Полагаме $z = \ln y, u = \ln x$ и получаваме $z = au + \ln c$.

В някои пакети приложни програми за корелационен анализ едновременно се изчислява статистическият коефициент на корелация и се построява корелационното поле за двойката случайни величини (X, Y). Предвидена е възможността тези процедури да се извършат и за някои двойки преобразувани случайни променливи като $(X, \dfrac{1}{Y})$, $(X, \log Y)$, $(\log X, Y)$, $(\log X, \log Y)$, $(\sqrt{X}, \log Y)$ и др. На двойката, чийто статистически коефициент на корелация има най-голяма абсолютна стойност, съответства най-силно изразена линейна зависимост. Така например, ако това е двойката $(\log X, Y)$, то съответното корелационно поле $(\log x_i, y_i)$ има най-ярко изразена емпирична линейна зависимост.

Таблица 16.1: Към пример 16.5

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
2	13	4	169	26
3	12	9	144	36
4	11	16	121	44
5	9	25	81	45
6	9	36	81	54
7	8	49	36	56
8	6	64	36	48
10	4	100	16	40
12	2	144	4	24
14	−1	196	1	−14
$S_x = 71$	$S_y = 73$	$S_{xx} = 643$	$S_{yy} = 717$	$S_{xy} = 359$

Пример 16.5. *В първите две колони на таблица* 16.1 *са дадени стойностите* (x_i, y_i), $i = 1, 2, \ldots, 10$ *на двумерната случайна величина* (X, Y). *Да се намери статистическият коефициент на корелация.*



Фиг. 16.2: $\overline{r}_{XY} = -0.9962$

Имаме, че

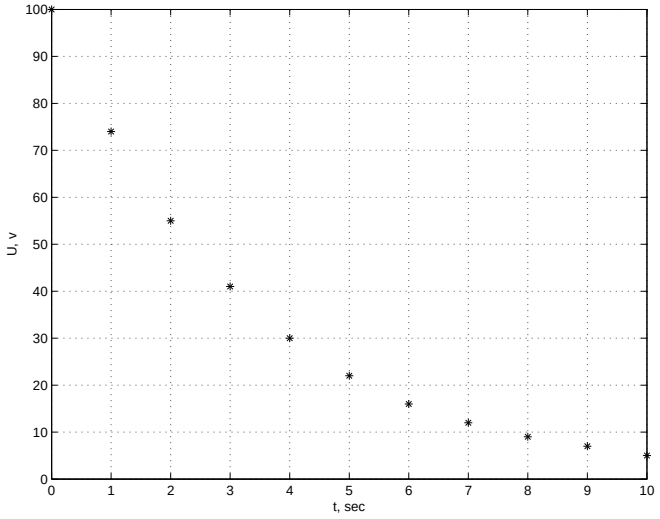
$$n = 10, \quad \overline{x} = \frac{S_x}{n} = \frac{71}{10} = 7.1, \quad \overline{y} = \frac{S_y}{n} = \frac{73}{10} = 7.3\,,$$
$$\overline{D}_X = \frac{nS_{xx} - S_x^2}{n^2} = \frac{10 \cdot 643 - 71^2}{100} = \frac{1389}{100} = 13.89\,,$$
$$\overline{D}_Y = \frac{nS_{yy} - S_y^2}{n^2} = \frac{10 \cdot 717 - 73^2}{100} = \frac{1841}{100} = 18.41\,,$$
$$\overline{K}_{XY} = \frac{nS_{xy} - S_x S_y}{n^2} = \frac{10 \cdot 359 - 71 \cdot 73}{100} = \frac{-1593}{100} = -15.93\,,$$
$$\overline{r}_{XY} = \frac{\overline{K}_{XY}}{\sqrt{\overline{D}_X \overline{D}_Y}} = \frac{-15.93}{\sqrt{13.89 \cdot 18.41}} = -0.9962.$$

Понеже коефициентът $\overline{r}_{XY}$ *е близък до* -1 *и точките на корелационното поле са групирани около права линия* (*фиг.* 16.2), *то предполагаме, че между величините* X *и* Y *има зависимост, която е близка до линейната и се представя с линейно уравнение на регресия* $y = a_0 + a_1 x$.

Пример 16.6. *В момент* $t = 0$ *кондензатор има напрежение* $U_0 = 100\,V$*, след което се разрежда през съпротивление. Напрежението* U *се измерва в различни моменти от времето* t *с точност* $1\,V$*. Резултатите са дадени в долната таблица.*

$X = t, s$	0	1	2	3	4	5	6	7	8	9	10
$Y = U, V$	100	74	55	41	30	22	16	12	9	7	5

Да се избере типа на функцията, която „пасва" на корелационното поле. След пресмятане получаваме $\overline{r}_{XY} = -0.927$*, т.е. имаме силна корелационна зависимост. Тази зависимост обаче не е линейна, което личи от корелационното поле (фиг.* 16.3*). Да пробваме дали зависимост-*



Фиг. 16.3: $\overline{r}_{XY} = -0.927$

та $U = be^{at}$ *не е по-подходяща. За целта преобразуваме променливите* $Y = \ln U$*,* $X = t$*. За новите променливи получаваме, че* $\overline{r}_{XY} = -0.9998$*. Понеже коефициентът* $\overline{r}_{XY}$ *е много по-близък до* -1*, може да предположим, че между променливите* $\ln U$ *и* t *има по-ярко изразена емпирична линейна зависимост, която се представя с линейно уравнение на регресия* $\ln U = a_0 + a_1 t$*.*

16.4 Регресионен анализ. Метод на най-малките квадрати

В регресионния анализ най-често се използва т.нар. **линеен регресионен модел**, в който неизвестните параметри a_i участват линейно в уравнението на регресия

$$y = a_0 + a_1\varphi_1(x) + a_2\varphi_2(x) + \cdots + a_m\varphi_m(x),$$

а фуннкциите $\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x)$ са известни и са избрани след направения корелационен анализ. Например при проста линейна регресия уравнението на регресия е линейно и има вида

$$y = a_0 + a_1x.$$

Ще отбележим, че към линейните модели спадат и тези с уравнение на регресия

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m.$$

В практическите задачи обикновено $m \leq 4$.

След избора на модела се преминава към определянето на параметрите така, че да получим линия на регресия, която „най-добре пасва" на точките от корелационното поле. Тези „оптимални" параметри се определят с метода на най-малките квадрати.

16.4.1 Метод на най-малките квадрати

Нека сме установили наличието на линейна корелация и сме избрали уравнението на линейна регресия

$$y = a_0 + a_1x. \tag{16.9}$$

Следващата задача е, като използваме извадката $(x_i, y_i),\ i = 1, 2, \cdots, n$, да намерим „оптимални" оценки $\overline{a}_0, \overline{a}_1$ за параметрите a_0, a_1. Това означава от всички прави с уравнение (16.9) да подберем тази, която е „най-близко" до точките (x_i, y_i).

За да оценим близостта на правата $y = a_0 + a_1x$ до точките (x_i, y_i), определяме числата

$$d_i^2 = (a_0 + a_1x_i - y_i)^2$$

– квадратите на отклоненията по вертикала (по y) на точките (x_i, y_i) от правата $y = a_0 + a_1 x$. Коефициентите a_0, a_1 се избират така, че сумата от всички такива квадрати да е минимална:

$$F = \sum_{i=1}^{n} (a_0 + a_1 x_i - y_i)^2 \to \min .$$

Понеже F е диференцуема функция на a_0 и a_1, то в точката на минимум задължително са изпълнени условията

$$\frac{\partial F}{\partial a_0} = 0, \quad \frac{\partial F}{\partial a_1} = 0 .$$

Имаме, че

$$\frac{\partial F}{\partial a_0} = 2 \sum_{i=1}^{n} (a_0 + a_1 x_i - y_i) = 0,$$
$$\frac{\partial F}{\partial a_1} = 2 \sum_{i=1}^{n} (a_0 + a_1 x_i - y_i) x_i = 0,$$

т.е.

$$a_0 \sum_{i=1}^{n} 1 + a_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 ,$$
$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0$$

и като отчетем (16.3) и (16.4), получаваме системата за a_0, a_1:

$$a_0 n + a_1 S_x = S_y, \tag{16.10}$$
$$a_0 S_x + a_1 S_{xx} = S_{xy}. \tag{16.11}$$

От (16.10) следва, че $\dfrac{S_y}{n} = a_0 + a_1 \dfrac{S_x}{n}$, т.е. $\overline{y} = a_0 + a_1 \overline{x}$, което означа, че точката $(\overline{x}, \overline{y})$ лежи върху правата $y = a_0 + a_1 x$. Заместваме $a_0 = \overline{y} - a_1 \overline{x}$ в (16.11) и като решим полученото уравнение, намираме

$$a_1 = \frac{\overline{K}_{XY}}{\overline{\sigma}_X^2} .$$

Окончателно, уравнението на линейна регресия е

$$y = \overline{a}_0 + \overline{a}_1 x, \tag{16.12}$$

където

$$\overline{a}_1 = \frac{\overline{K}_{XY}}{\overline{\sigma}_X^2}, \quad \overline{a}_0 = \overline{y} - \overline{a}_1 \overline{x}. \tag{16.13}$$

Нека търсим линейна регресия от вида $x = c_0 + c_1 y$. Прилагаме метода на най-малките квадрати, като вместо квадратите на отклоненията по y вземем квадратите на отклоненията по x : $q_i^2 = (c_0 + c_1 y_i - x_i)^2$. Тогава аналогично получаваме, че уравнението на линейна регресия е

$$x = \overline{c}_0 + \overline{c}_1 y, \tag{16.14}$$

където

$$\overline{c}_1 = \frac{\overline{K}_{XY}}{\overline{\sigma}_Y^2}, \quad \overline{c}_0 = \overline{x} - \overline{c}_1 \overline{y}. \tag{16.15}$$

Пример 16.7. *Да се намерят правите на регресия за величините* (X, Y) *от пример* 16.5. *Вече пресметнахме, че*

$$\overline{x} = 7.1, \quad \overline{y} = 7.3\,,$$
$$\overline{\sigma}_X^2 = \overline{D}_X = 13.89,\ \overline{\sigma}_Y^2 = \overline{D}_Y = 18.41,\ \overline{K}_{XY} = -15.93.$$

Тогава от (16.13) *и* (16.15) *следва,че*

$$\overline{a}_1 = \frac{\overline{K}_{XY}}{\overline{\sigma}_X^2} = \frac{-15.93}{13.89} \approx -1.1469\,,$$
$$\overline{a}_0 = \overline{y} - \overline{a}_1 \overline{x} = 7.3 - (-1.469) \cdot 7.1 \approx 15.443,$$
$$\overline{c}_1 = \frac{\overline{K}_{XY}}{\overline{\sigma}_Y^2} = \frac{-15.93}{18.41} \approx -0.8653\,,$$
$$\overline{c}_0 = \overline{x} - \overline{c}_1 \overline{y} = 7.1 - (-0.865) \cdot 7.3 \approx 13.417.$$

и съгласно (16.12) *и* (16.14) *уравненията на линейна регресия имат вида*

$$y = -1.1469x + 15.443, \quad x = -0.8653y + 13.417.$$

Правата на регресия $y = -1.1469x + 15.443$ *е дадена на фиг.* 16.2.

Когато уравнението на регресия има вида

$$y = g(x; a_0, a_1, \ldots, a_m),$$

оптималните оценки за параметрите $a_0, a_1, \ldots, a_m$ се получават, като се приложи методът на най-малките квадрати, т.е. като се минимизира функцията

$$F = \sum_{i=1}^{n} \big(g(x_i; a_0, a_1, \ldots, a_m) - y_i\big)^2 \to \min .$$

Решенията $\overline{a}_0, \overline{a}_1, \ldots, \overline{a}_m$ на тази задача се получават, като се реши системата

$$\frac{\partial F}{\partial a_0} = 0, \quad \frac{\partial F}{\partial a_1} = 0, \ldots, \frac{\partial F}{\partial a_m} = 0 \cdot \tag{16.16}$$

Когато уравнението на регресия

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m$$

е линейно относно параметрите a_i, получената система (16.16) е също линейна и може да бъде решена относно $a_0, a_1, \ldots, a_m$.

Задача 16.1. *Като се приложи методът на най-малките квадрати, да се намери системата за определяне на оптималните оценки* $\overline{a}_0, \overline{a}_1, \overline{a}_2$ *на параметрите* a_0, a_1, a_2*, ако уравнението на регресия има вида*

$$y = a_0 + a_1 x + a_2 x^2.$$

16.4.2 Интервални оценки на параметрите

Когато имаме линейна регресия от вида $y = a_0 + a_1 x$, резултатите от наблюденията $(x_i, y_i),\ i = 1, 2, \ldots, n$ може да се представят във вида

$$y_i = a_0 + a_1 x_i + e_i,$$

където остатъците e_i са случайните грешки от наблюдението.

Предполагаме, че грешките от наблюдението e_i имат нулеви математически очаквания, равни дисперсии σ_e^2 и не са корелирани, т.е.

$$Ee_i = 0,\ De_i = \sigma_e^2, \tag{16.17}$$

$$K_{e_i e_j} = 0,\ i \neq j. \tag{16.18}$$

Нека $\overline{a}_0, \overline{a}_1$ са точковите оценки за параметрите a_0, a_1, получени по метода на най-малките квадрати (формули (16.13)). Тогава сумата от квадратите на грешките

$$Q_e = \sum_{i=1}^{n} e_i^2$$

е минимална. Може да се докаже, че

$$Q_e = n(\overline{D}_Y - \overline{a}_1 \overline{K}_{XY}), \tag{16.19}$$

където $\overline{D}_Y, \overline{K}_{XY}$ и $\overline{a}_1$ са пресметнати по формулите (16.6),(16.7) и (16.13).

Неизместената оценка на дисперсията σ_e^2 на грешките от наблюденията е равна на

$$\overline{s}_e^2 = \frac{Q_e}{n-2} \cdot \tag{16.20}$$

При предположение, че грешките от наблюденията e_i са некорелирани и имат нормално разпределение $N(0, \sigma_e)$, се определят границите на доверителните интервали за параметрите a_0, a_1 при зададено ниво на значимост α или доверителна вероятност $p = 1 - \alpha$:

$$a_0 \in (\overline{a}_0 - \Delta_0, \overline{a}_0 + \Delta_0), \quad a_1 \in (\overline{a}_1 - \Delta_1, \overline{a}_1 + \Delta_1). \tag{16.21}$$

Тук

$$\Delta_0 = t_{1-\alpha/2}(n-2)\sqrt{\frac{\overline{s}_e^2 S_{xx}}{n^2 \overline{D}_X}}, \quad \Delta_1 = t_{1-\alpha/2}(n-2)\sqrt{\frac{\overline{s}_e^2}{n\overline{D}_X}}, \tag{16.22}$$

S_{xx} е пресметнато по формула (16.4), а квантилът $t_{1-\alpha/2}(n-2)$ на разпределението на Стюдънт $T(n-2)$ с $(n-2)$ степени на свобода се определя от Таблица 6.

16.4.3 Проверка за съгласуваност

Проверката за съгласуваност на линейната регресия с резултатите от наблюденията се свежда до проверка на хипотезата за липса на линейна връзка между променливите X и Y, т.е. хипотезата $H_0 : a_1 = 0$.

Хипотезата H_0 се приема при ниво на значимост α, ако доверителният интервал за параметъра a_1 съдържа нулата: $0 \in (\overline{a}_1 - \Delta_1, \overline{a}_1 + \Delta_1)$. В този случай се счита, че линейната регресия не се съгласува с резултатите от

наблюденията и за представяне на данните трябва да се използва друг модел (например модел от втори ред $y = a_0 + a_1x + a_2x^2$).

За проверка на хипотезата $H_0 : a_1 = 0$ също може да се използва статистиката

$$F = \frac{n\overline{a}_1\overline{K}_{XY}}{\overline{s}_e^2}\,.$$

Ако хипотезата H_0 е вярна, тази статистика има разпределение на Фишер $F(1, n-2)$. Ако F_0 е наблюдаваната стойност на F и

$$F_0 < F_{1-\alpha}(1, n-2),$$

то хипотезата H_0 се приема.

Пример 16.8. *В таблица* 16.1 *са дадени* $n = 10$ *наблюдения на величините* X, Y. *В пример* 16.5 *са намерени оценките*

$\overline{D}_X = 13.89$, $\overline{D}_Y = 18.41$, $\overline{K}_{XY} = -15.93$ *и стойността* $S_{xx} = 643$.

В пример 16.7, *по метода на най-малките квадрати са намерени оценките* $\overline{a}_0 = 15.443$ *и* $\overline{a}_1 = -1.1469$ *на параметрите* a_0, a_1 *на линията на линейна регресия* $y = a_0 + a_1x$.

С ниво на значимост $\alpha = 0.05$ *да се определят доверителните интервали* I_0 *и* I_1 *за* a_0 *и* a_1 *и да се направи проверка за съгласуваност на линейната регресия с резултатите от наблюдението. Предполага се, че грешките от наблюдението са независими и имат нормално разпределение* $N(0, \sigma_e)$.

Решение: По формулите (16.19) *и* (16.20) *пресмятаме, че*

$$Q_e = 10(18.41 - (-1.1469)\cdot(-15.93)) \approx 1.4,\ \ \overline{s}_e^2 = \frac{Q_e}{n-2} = \frac{1.4}{8} = 0.175.$$

От Таблица 6 *определяме квантила* $t_{1-\alpha/2}(n-2) = t_{0.975}(8) = 2.131$ *и по формулите* (16.22) *намираме:*

$$\Delta_0 = 2.131\sqrt{\frac{0.175\cdot 643}{100\cdot 13.89}} \approx 0.6065,\ \ \Delta_1 = 2.131\sqrt{\frac{0.175}{10\cdot 13.89}} \approx 0.0756.$$

Тогава съгласно (16.21) *доверителните интервали за* a_0 *и* a_1 *са:*

$$I_0 = (14.8365, 16.0495),\ \ I_1 = (-1.2225, -1.0713).$$

Понеже $0 \notin I_1$, *то хипотезата* $H_0 : a_1 = 0$ *се отхвърля; следователно линейната регресия се съгласува с резултатите от наблюдението.*