Основы регрессионного анализа Исходные предположения. Метод наименьших квадратов

1. Линейная модель парной регрессии и корреляции

Рассмотрим простейшую модель парной регрессии – линейную регрессию. Линейная регрессия находит широкое применение в эконометрике ввиду четкой экономической интерпретации ее параметров.

Линейная регрессия сводится к нахождению уравнения вида:

$$\hat{y}_x = a + b \cdot x \text{ или } y = a + b \cdot x + \varepsilon. \quad (1.1)$$

Уравнение вида $\hat{y}_x = a + b \cdot x$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x.

Построение линейной регрессии сводится к оценке ее параметров – a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических $\hat{y}_x$ минимальна:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_{x_i}\right)^2 = \sum_{i=1}^{n}\varepsilon_i^2 \rightarrow \min. \quad (1.2)$$

Т. е. из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной (рис. 1.2):

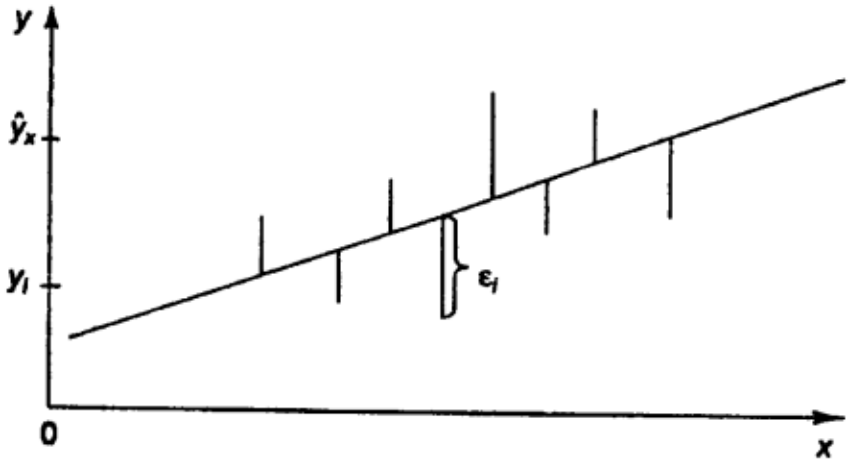

Рис. 1.2. Линия регрессии с минимальной дисперсией остатков

Как известно из курса математического анализа, чтобы найти минимум функции (1.2), надо вычислить частные производные по каждому из параметров a и b и приравнять их к нулю. Обозначим $\sum_i \varepsilon_i^2$ через $S(a, b)$, тогда:

$$S(a, b) = \sum (y - a - b \cdot x)^2.$$

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum (y - a - b \cdot x) = 0; \\ \dfrac{\partial S}{\partial b} = -2\sum x(y - a - b \cdot x) = 0. \end{cases} \quad (1.3)$$

После несложных преобразований получим следующую систему линейных уравнений для оценки параметров a и b:

$$\begin{cases} a \cdot n + b \cdot \sum x = \sum y; \\ a \cdot \sum x + b \cdot \sum x^2 = \sum x \cdot y. \end{cases} \quad (1.4)$$

Решая систему уравнений (1.4), найдем искомые оценки параметров a и b. Можно воспользоваться следующими готовыми формулами, которые следуют непосредственно из решения системы (1.4):

$$a = \overline{y} - b \cdot \overline{x}, \qquad b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}, \quad (1.5)$$

где $\operatorname{cov}(x, y) = \overline{y \cdot x} - \overline{y} \cdot \overline{x}$ – ковариация признаков x и y, $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ – дисперсия признака x и

$$\overline{x} = \frac{1}{n}\sum x, \ \overline{y} = \frac{1}{n}\sum y, \ \overline{y \cdot x} = \frac{1}{n}\sum y \cdot x, \ \overline{x^2} = \frac{1}{n}\sum x^2.$$

Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий. Дисперсия – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания. Математическое ожидание – сумма произведений значений случайной величины на соответствующие вероятности[2].

Параметр b называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу.

Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях.

Формально a – значение y при $x = 0$. Если признак-фактор x не может иметь нулевого значения, то вышеуказанная трактовка свободного члена a не имеет смысла, т. е. параметр a может не иметь экономического содержания.

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает линейный коэффициент корреляции r_{xy}, который можно рассчитать по следующим формулам:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y}. \qquad (1.6)$$

Линейный коэффициент корреляции находится в пределах: $-1 \le r_{xy} \le 1$. Чем ближе абсолютное значение r_{xy} к единице, тем сильнее линейная связь между факторами (при $r_{xy} = \pm 1$ имеем строгую функциональную зависимость). Но следует иметь в виду, что близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствия связи между признаками. При другой (нелинейной) спецификации модели связь между признаками может оказаться достаточно тесной.

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r_{xy}^2, называемый коэффициентом детерминации. Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака:

$$r_{xy}^2 = 1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}, \qquad (1.7)$$

где $\sigma^2_{\text{ост}} = \frac{1}{n}\sum\left(y - \hat{y}_x\right)^2$, $\sigma^2_y = \frac{1}{n}\sum\left(y - \overline{y}\right)^2 = \overline{y^2} - \overline{y}^2$.

Соответственно величина $1 - r^2_{xy}$ характеризует долю дисперсии y, вызванную влиянием остальных, не учтенных в модели факторов.

После того как найдено уравнение линейной регрессии проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации:

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y - \hat{y}_x}{y}\right| \cdot 100\%. \qquad (1.8)$$

Средняя ошибка аппроксимации не должна превышать 8–10%.

Оценка значимости уравнения регрессии в целом производится на основе F-критерия Фишера, которому предшествует дисперсионный анализ. В математической статистике дисперсионный анализ рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели.

Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной y от среднего значения $\overline{y}$ раскладывается на две части – «объясненную» и «необъясненную»:

$$\sum\left(y - \overline{y}\right)^2 = \sum\left(\hat{y}_x - \overline{y}\right)^2 + \sum\left(y - \hat{y}_x\right)^2,$$

где $\sum\left(y - \overline{y}\right)^2$ – общая сумма квадратов отклонений; $\sum\left(\hat{y}_x - \overline{y}\right)^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений); $\sum\left(y - \hat{y}_x\right)^2$ – остаточная сумма квадратов

отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в таблице 1.1 (n – число наблюдений, m – число параметров при переменной x).

Таблица 1.1

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum\left(y-\overline{y}\right)^2$	$n-1$	$S^2_{\text{общ}}=\dfrac{\sum\left(y-\overline{y}\right)^2}{n-1}$
Факторная	$\sum\left(\hat{y}_x-\overline{y}\right)^2$	m	$S^2_{\text{факт}}=\dfrac{\sum\left(\hat{y}_x-\overline{y}\right)^2}{m}$
Остаточная	$\sum\left(y-\hat{y}_x\right)^2$	$n-m-1$	$S^2_{\text{ост}}=\dfrac{\sum\left(y-\hat{y}_x\right)^2}{n-m-1}$

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера:

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}. \tag{1.9}$$

Фактическое значение F-критерия Фишера (1.9) сравнивается с табличным значением $F_{\text{табл}}\left(\alpha;k_1;k_2\right)$ при уровне значимости α и степенях свободы $k_1=m$ и $k_2=n-m-1$. При этом, если фактическое значение F-критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для парной линейной регрессии $m=1$, поэтому:

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}=\frac{\sum\left(\hat{y}_x-\overline{y}\right)^2}{\sum\left(y-\hat{y}_x\right)^2}\cdot\left(n-2\right). \tag{1.10}$$

Величина F-критерия связана с коэффициентом детерминации r^2_{xy}, и ее можно рассчитать по следующей формуле:

$$F=\frac{r^2_{xy}}{1-r^2_{xy}}\cdot\left(n-2\right). \tag{1.11}$$

В парной линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров определяется его стандартная ошибка: m_b и m_a.

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \sqrt{\frac{S^2_{\text{ост}}}{\sum\left(x-\overline{x}\right)^2}} = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}}, \qquad (1.12)$$

где $S^2_{\text{ост}} = \dfrac{\sum\left(y-\hat{y}_x\right)^2}{n-2}$ – остаточная дисперсия на одну степень свободы.

Величина стандартной ошибки совместно с t-распределением Стьюдента при $n-2$ степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительного интервала.

Для оценки существенности коэффициента регрессии его величина сравнивается с его стандартной ошибкой, т. е. определяется фактическое значение t-критерия Стьюдента: $t_b = \dfrac{b}{m_b}$, которое затем сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы $(n-2)$. Доверительный интервал для коэффициента регрессии определяется как $b \pm t_{\text{табл}} \cdot m_b$. Поскольку знак коэффициента регрессии указывает на рост результативного признака y при увеличении признака-фактора x $(b>0)$, уменьшение результативного признака при увеличении признака-фактора $(b<0)$ или его независимость от независимой переменной $(b=0)$ (см. рис. 1.3), то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например $-1{,}5 \le b \le 0{,}8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

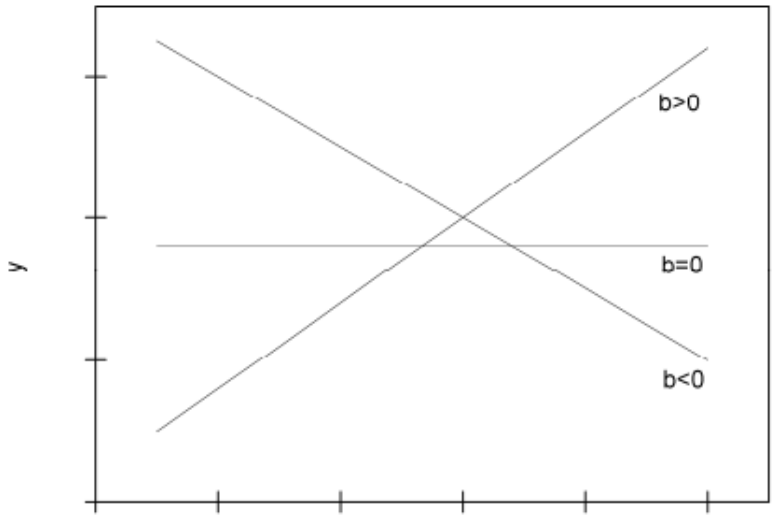

Рис. 1.3. Наклон линии регрессии в зависимости от значения параметра b

Стандартная ошибка параметра a определяется по формуле:

$$m_a = \sqrt{S_{\text{ост}}^2 \cdot \frac{\sum x^2}{n \cdot \sum (x - \bar{x})^2}} = S_{\text{ост}} \cdot \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n}. \qquad (1.13)$$

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии. Вычисляется t-критерий: $t_a = \frac{a}{m_a}$, его величина сравнивается с табличным значением при $n-2$ степенях свободы.

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции m_r:

$$m_r = \sqrt{\frac{1-r^2}{n-2}}. \qquad (1.14)$$

Фактическое значение t-критерия Стьюдента определяется как $t_r = \frac{r}{m_r}$.

Существует связь между t-критерием Стьюдента и F-критерием Фишера:

$$t_b = t_r = \sqrt{F}. \qquad (1.15)$$

В прогнозных расчетах по уравнению регрессии определяется предсказываемое $\hat{y}_p$ значение как точечный прогноз $\hat{y}_x$ при $x_p = x_k$, т. е.

путем подстановки в уравнение регрессии $\hat{y}_x = a + b \cdot x$ соответствующего значения x. Однако точечный прогноз явно не реален. Поэтому он дополняется расчетом стандартной ошибки $\hat{y}_p$, т.е. $m_{\hat{y}_p}$, и соответственно интервальной оценкой прогнозного значения $\hat{y}_p$:

$$\hat{y}_p - \Delta_{\hat{y}_p} \le \hat{y}_p \le \hat{y}_p + \Delta_{\hat{y}_p},$$

где $\Delta_{\hat{y}_p} = m_{\hat{y}_p} \cdot t_{\text{табл}}$, а $m_{\hat{y}_p}$ – средняя ошибка прогнозируемого индивидуального значения:

$$m_{\hat{y}_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{n \cdot \sigma_x^2}}. \qquad (1.16)$$

Рассмотрим **пример**. По данным проведенного опроса восьми групп семей известны данные связи расходов населения на продукты питания с уровнем доходов семьи.

Таблица 1.2

Расходы на продукты питания, y, тыс. руб.	0, 9	1, 2	1, 8	2, 2	2, 6	2, 9	3, 3	3, 8
Доходы семьи, x, тыс. руб.	1, 2	3, 1	5, 3	7, 4	9, 6	11, 8	14, 5	18, 7

Предположим, что связь между доходами семьи и расходами на продукты питания линейная. Для подтверждения нашего предположения построим поле корреляции.

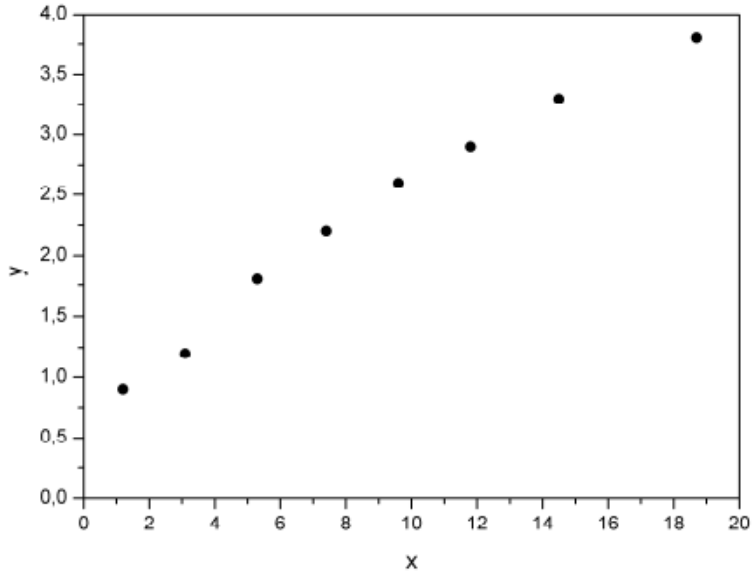

Рис. 1.4

По графику видно, что точки выстраиваются в некоторую прямую линию.

Для удобства дальнейших вычислений составим таблицу.

Таблица 1.3

	x	y	$x \cdot y$	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	$\left(y - \hat{y}_x\right)^2$	A_i, %
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1, 2	0, 9	1,08	1,44	0,81	1,038	–0,138	0,0190	15,33
2	3, 1	1, 2	3,72	9,61	1,44	1,357	–0,157	0,0246	13,08
3	5, 3	1, 8	9,54	28,09	3,24	1,726	0,074	0,0055	4,11
4	7, 4	2, 2	16,28	54,76	4,84	2,079	0,121	0,0146	5,50
5	9, 6	2, 6	24,96	92,16	6,76	2,449	0,151	0,0228	5,81
6	11, 8	2, 9	34,22	139,24	8,41	2,818	0,082	0,0067	2,83
7	14, 5	3, 3	47,85	210,25	10,89	3,272	0,028	0,0008	0,85

Окончание табл. 1.3.

	x	y	$x \cdot y$	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	$\left(y - \hat{y}_x\right)^2$	A_i, %
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
8	18,7	3,8	71,06	349,69	14,44	3,978	–0,178	0,0317	4,68
Итого	71,6	18,7	208,71	885,24	50,83	18,717	–0,017	0,1257	52,19
Среднее значение	8,95	2,34	26,09	110,66	6,35	2,34	–	0,0157	6,52
σ	5,53	0,935	–	–	–	–	–	–	–
σ^2	30,56	0,874	–	–	–	–	–	–	–

Рассчитаем параметры линейного уравнения парной регрессии $\hat{y}_x = a + b \cdot x$. Для этого воспользуемся формулами (1.5):

$$b = \frac{\operatorname{cov}(x,\, y)}{\sigma_x^2} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{26,09 - 8,95 \cdot 2,34}{30,56} = 0,168;$$

$$a = \overline{y} - b \cdot \overline{x} = 2,34 - 0,168 \cdot 8,95 = 0,836.$$

Получили уравнение: $\hat{y}_x = 0,836 + 0,168 \cdot x$. Т. е. с увеличением дохода семьи на 1000 руб. расходы на питание увеличиваются на 168 руб.

Как было указано выше, уравнение линейной регрессии всегда дополняется показателем тесноты связи – линейным коэффициентом корреляции r_{xy}:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0,168 \cdot \frac{5,53}{0,935} = 0,994.$$

Близость коэффициента корреляции к 1 указывает на тесную линейную связь между признаками.

Коэффициент детерминации $r_{xy}^2 = 0,987$ (примерно тот же результат получим, если воспользуемся формулой (1.7)) показывает, что уравнением регрессии объясняется 98, 7% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1, 3%.

Оценим качество уравнения регрессии в целом с помощью F-критерия Фишера. Сосчитаем фактическое значение F-критерия:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2) = \frac{0,987}{1-0,987} \cdot 6 = 455,54.$$

Табличное значение ($k_1 = 1$, $k_2 = n - 2 = 6$, $\alpha = 0,05$): $F_{\text{табл}} = 5,99$. Так как $F_{\text{факт}} > F_{\text{табл}}$, то признается статистическая значимость уравнения в целом.

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитаем t-критерий Стьюдента и доверительные интервалы каждого из показателей. Рассчитаем случайные ошибки параметров линейной регрессии и коэффициента корреляции $\left(S_{\text{ост}}^2 = \frac{\sum \left(y - \hat{y}_x \right)^2}{n-2} = \frac{0,1257}{8-2} = 0,021 \right)$:

$$m_b = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}} = \frac{\sqrt{0,021}}{5,53 \cdot \sqrt{8}} = 0,0093,$$

$$m_a = S_{\text{ост}} \cdot \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n} = \frac{\sqrt{0,021 \cdot 885,24}}{5,53 \cdot 8} = 0,0975,$$

$$m_r = \sqrt{\frac{1 - r^2}{n-2}} = \sqrt{\frac{1-0,987}{6}} = 0,0465.$$

Фактические значения t-статистик: $t_b = \frac{0,168}{0,0093} = 18,065$, $t_a = \frac{0,836}{0,0975} = 8,574$, $t_r = \frac{0,994}{0,0465} = 21,376$. Табличное значение t-

критерия Стьюдента при $\alpha = 0,05$ и числе степеней свободы $\nu = n - 2 = 6$ есть $t_{\text{табл}} = 2,447$. Так как $t_b > t_{\text{табл}}$, $t_a > t_{\text{табл}}$ и $t_r > t_{\text{табл}}$, то признаем статистическую значимость параметров регрессии и показателя тесноты связи. Рассчитаем доверительные интервалы для параметров регрессии a и b: $a \pm t \cdot m_a$ и $b \pm t \cdot m_b$. Получим, что $a \in [0,597;\ 1,075]$ и $b \in [0,145;\ 0,191]$.

Средняя ошибка аппроксимации (находим с помощью столбца 10 таблицы 1.3; $A_i = \left| \frac{y_i - \hat{y}_{x_i}}{y_i} \right| \cdot 100\%$) $\overline{A} = 6,52\%$ говорит о хорошем качестве уравнения регрессии, т. е. свидетельствует о хорошем подборе модели к исходным данным.

И, наконец, найдем прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110 % от среднего уровня $x_p = 1,1 \cdot \overline{x} = 1,1 \cdot 8,95 = 9,845$, т. е. найдем расходы на питание, если доходы семьи составят 9,85 тыс. руб.:

$$\hat{y}_p = 0,836 + 0,168 \cdot 9,845 = 2,490 \text{ (тыс. руб.)}.$$

Значит, если доходы семьи составят 9,845 тыс. руб., то расходы на питание будут 2,490 тыс. руб.

Найдем доверительный интервал прогноза. Ошибка прогноза:

$$m_{\hat{y}_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{n \cdot \sigma_x^2}} = \sqrt{0,021 \cdot \left(1 + \frac{1}{8} + \frac{\left(9,845 - 8,95\right)^2}{8 \cdot 30,56}\right)} = 0,154,$$

а доверительный интервал ($\hat{y}_p - \Delta_{\hat{y}_p} \le \hat{y}_p \le \hat{y}_p + \Delta_{\hat{y}_p}$):

$$2,113 < \hat{y}_p < 2,867.$$

Т. е. прогноз является статистически надежным.

Теперь на одном графике изобразим исходные данные и линию регрессии:

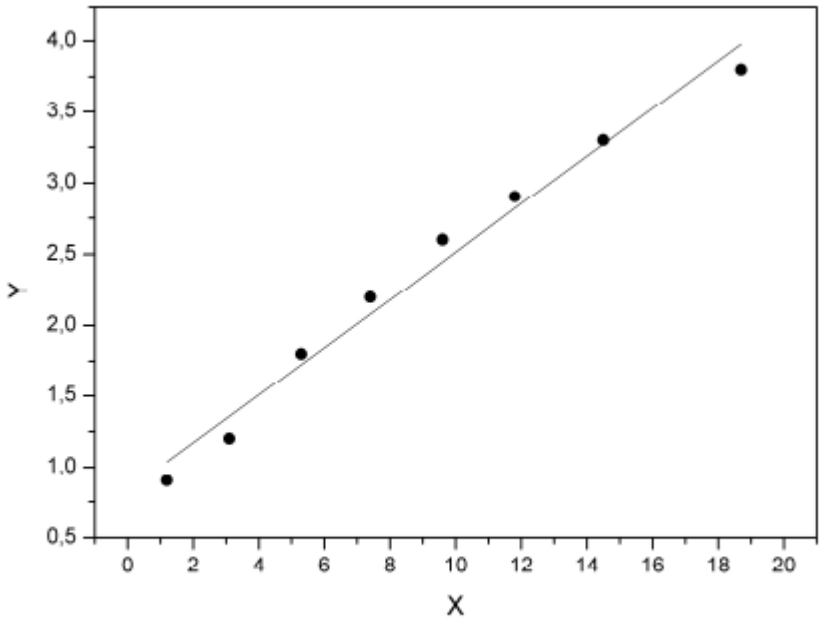

Рис. 1.5