# Эконометрика


## Кракашова Ольга Анатольевна

канд. экон. наук,
доцент кафедры СЭиОР РГЭУ (РИНХ)

# Лекция № 2

# Парная (простая) регрессия

# Парная регрессия и корреляция

Парная (простая) регрессия представляет собой модель, где среднее значение зависимой (объясняемой) переменной рассматривается как функция одной независимой (объясняющей) переменной x , т.е. это модель вида:

$$\hat{y}_x = f(x).$$

В каждом отдельном случае величина *y складывается* из двух слагаемых:

$$y = \hat{y}_x + \varepsilon,$$

где *y – фактическое значение результативного признака;* $\hat{y}_x$ - теоретическое значение результативного признака, найденное исходя из уравнения регрессии; ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии.

Случайная величина ε называется также возмущением. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

# Виды ошибок при построении регрессии и методы их устранения

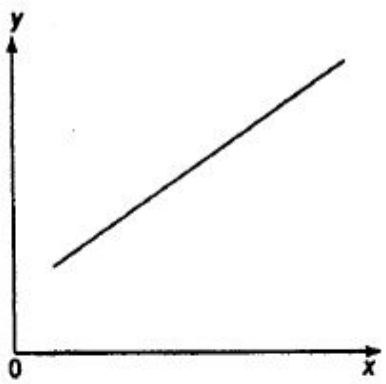
| Вид ошибки | Метод устранения |
|---|---|
| Ошибки спецификации: неправильный выбор той или иной математической функции для *и недоучет в уравнении регрессии* какого-либо существенного фактора, т.е. использование парной регрессии вместо множественной. | Изменение формы модели (вид математической формулы). |
| Ошибки выборки, которые имеют место в силу неоднородности данных в исходной статистической совокупности. | Увеличение объема исходных данных; исключение из совокупности единицы с аномальными значениями исследуемых признаков. |
| Ошибки измерения | Изменение методики измерения (когда это возможно). |

Предполагая, что ошибки измерения сведены к минимуму, основное внимание в эконометрических исследованиях уделяется ошибкам спецификации модели.

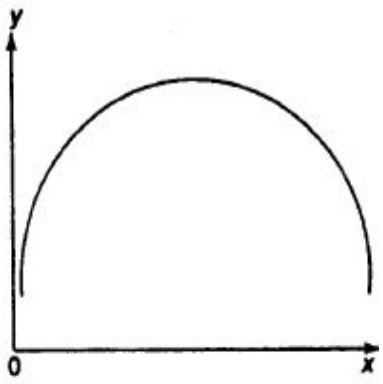
**Методы выбора вида математической функции:**

1) графический;
2) аналитический, т.е. исходя из теории изучаемой взаимосвязи;
3) экспериментальный.

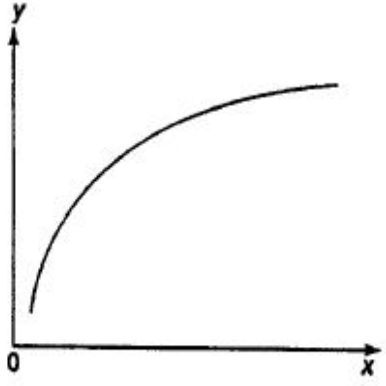
# Основные типы кривых, используемые при количественной оценке связей ме жду двумя переменными



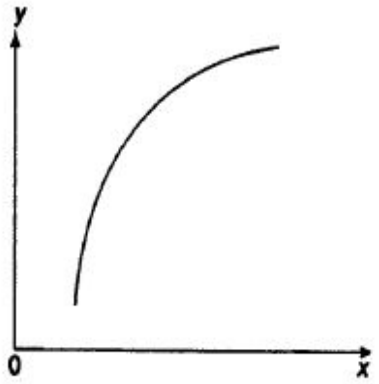

$\hat{y}_x = a + b \cdot x$



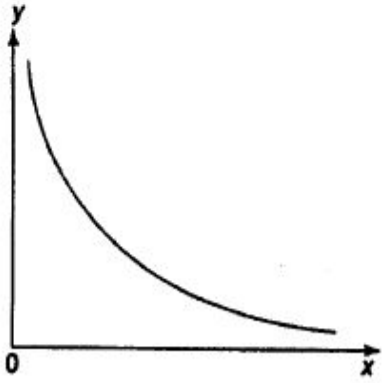

$\hat{y}_x = a + b \cdot x + c \cdot x^2$



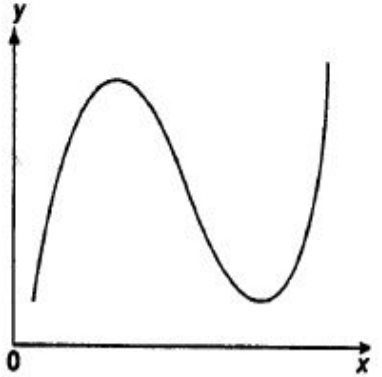

$\hat{y}_x = a \cdot x^b$




$\hat{y}_x = a \cdot b^x$




$\hat{y}_x = a + b/x$




$\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$

Если уравнение регрессии проходит через все точки корреляционного поля, что возможно только при функциональной связи, когда все точки лежат на линии регрессии $\hat{y}_x = f(x)$, то фактические значения результативного признака совпадают с теоретическими $y = \hat{y}_x$, т.е. они полностью обусловлены влиянием фактора $x$. В этом случае остаточная дисперсия $\sigma^2_{ост} = 0$.

В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии, факторов. Иными словами, имеют место отклонения фактических данных от теоретических $\left(y - \hat{y}_x\right)$. Величина этих отклонений и лежит в основе расчета остаточной дисперсии:

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

Чем меньше величина остаточной дисперсии, тем меньше влияние не учитываемых в уравнении регрессии факторов и тем лучше уравнение регрессии подходит к исходным данным.

Считается, что число наблюдений должно в 7-8 раз превышать число рассчитываемых параметров при переменной $x$. Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений, ибо каждый параметр при $x$ должен рассчитываться хотя бы по 7 наблюдениям. Значит, если мы выбираем параболу второй степени $\hat{y}_x = a + b \cdot x + c \cdot x^2$, то требуется объем информации уже не менее 14 наблюдений.

# Линейная модель парной регрессии и корреляции

Линейная регрессия сводится к нахождению уравнения вида

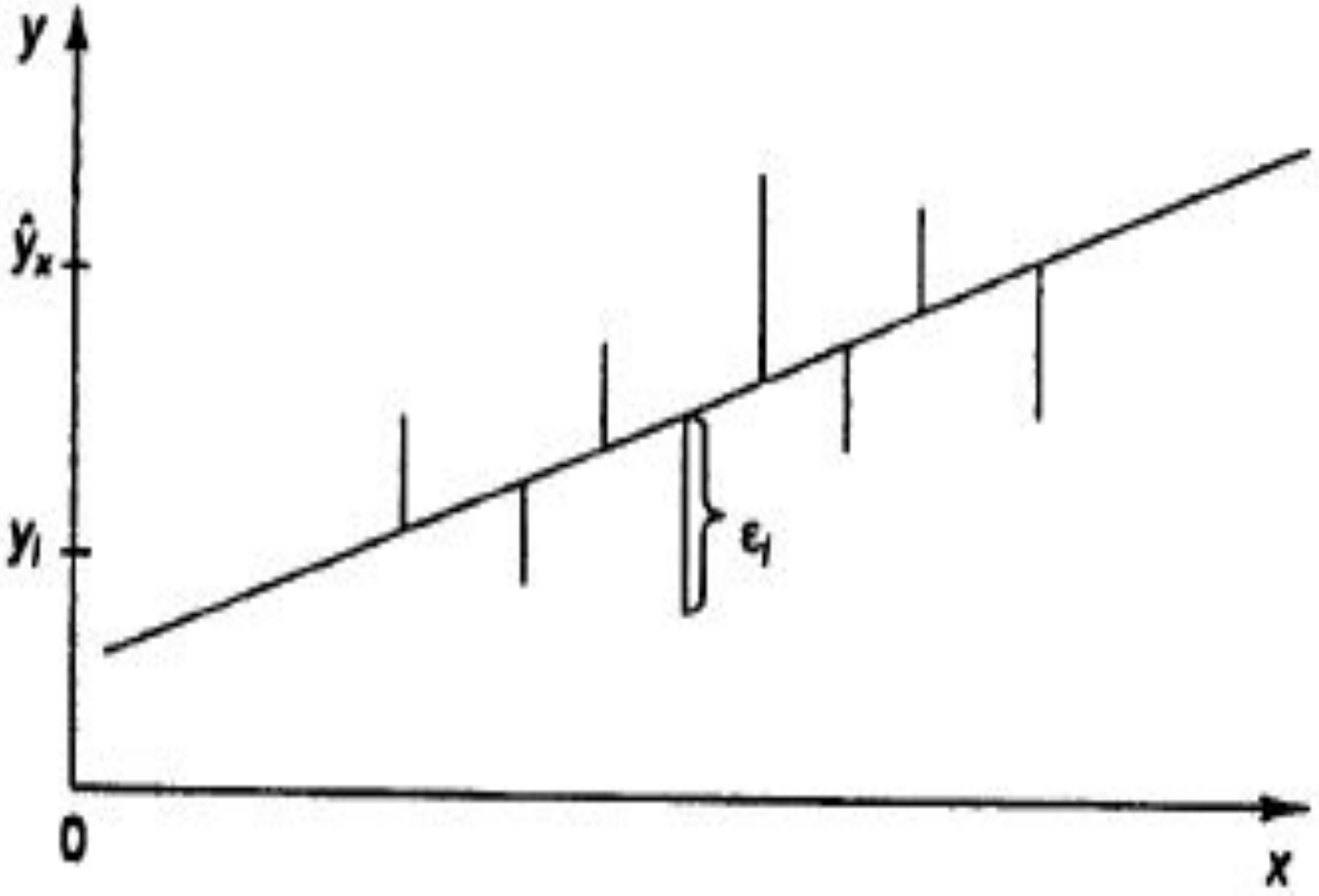
$$\hat{y}_x = a + b \cdot x \text{ или } y = a + b \cdot x + \varepsilon. \quad (1.1)$$

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:

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

# Линия регрессии с минимальной дисперсией остатков



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

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

**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.

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

# Показатель тесноты связи при использовании линейной регрессии - линейный коэффициент корреляции:

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

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.

# Коэффициент детерминации

Коэффициент детерминации характеризует долю дисперсии результативного признака $y$, объясняемую регрессией, в общей дисперсии результативного признака:

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

где $\sigma_y^2 = \frac{1}{n}\sum(y-\bar{y})^2$, $\sigma_{\text{факт}}^2 = \frac{1}{n}\sum(\hat{y}_x-\bar{y})^2$, $\sigma_{\text{ост}}^2 = \frac{1}{n}\sum(y-\hat{y}_x)^2$.

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

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.

# Средняя ошибка 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.

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

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

где $\sum(y-\overline{y})^2$ – 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

## (*n – число наблюдений, m – число параметров при переменной x*)

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

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

$$F=\frac{S^2_{факт}}{S^2_{ост}}. \qquad (1.9)$$

Фактическое значение $F$-критерия Фишера (1.9) сравнивается с табличным значением $F_{\text{табл}}(\alpha; k_1; k_2)$ при уровне значимости $\alpha$ 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

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

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

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

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:

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

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:*

$$t_b = \frac{b}{m_b}$$

которое затем сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы (*n - 2). Доверительный интервал для коэффициента* регрессии определяется как . $\overline{b} \pm t_{табл} \cdot m_b$ *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), то* границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например, -1,5 ≤ *b ≤ 0,8. Такого рода запись указывает, что истинное значение* коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

Стандартная ошибка параметра *a* определяется по формуле:

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

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:

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

он дополняется расчетом стандартной ошибки

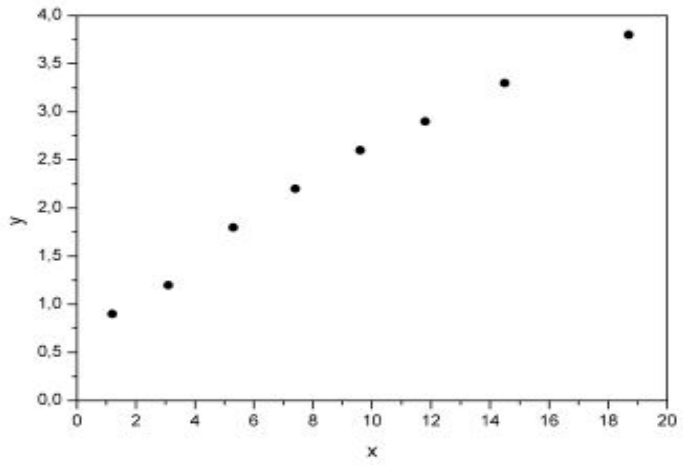
$$m_{\hat{y}_0} = \sqrt{S^2_{\text{ост}}\left(1+\frac{1}{n}+\frac{(x_p-\overline{x})^2}{\sum(x-\overline{x})^2}\right)} = \sqrt{S^2_{\text{ост}}\left(1+\frac{1}{n}+\frac{(x_p-\overline{x})^2}{n\cdot\sigma_x^2}\right)}, \qquad (1.16)$$

где $S^2_{\text{ост}} = \frac{\sum(y-\hat{y}_x)^2}{n-2}$, и построением доверительного интервала прогнозного значения $y_0^*$:

$$\hat{y}_0 - m_{\hat{y}_0}\cdot t_{\text{табл}} \le y_0^* \le \hat{y}_x + m_{\hat{y}_x}\cdot t_{\text{табл}}$$

# Пример

| 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 |




Поле корреляции.

| | $x$ | $y$ | $x \cdot y$ | $x^2$ | $y^2$ | $\hat{y}_x$ | $y-\hat{y}_x$ | $(y-\hat{y}_x)^2$ | $A_i$, % |
|---|---|---|---|---|---|---|---|---|---|
| **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 |
| 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 |
| $\sigma$ | 5,53 | 0,935 | – | – | – | – | – | – | – |
| $\sigma^2$ | 30,56 | 0,874 | – | – | – | – | – | – | – |

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

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

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 $r_{xy}$:

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

$$r_{xy}^2 = 0,987$$

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 $(k_1 = 1, \quad k_2 = n-2 = 6, \quad \alpha = 0,05)$: $F_{\text{табл}} = 5,99$. 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 $\hat{y}_0$ при значении признака-фактора, составляющем 110% от среднего уровня $x_0 = 1,1 \cdot \overline{x} = 1,1 \cdot 8,95 = 9,845$, т.е. найдем расходы на питание, если доходы семьи составят 9,85 тыс. руб.

$$\hat{y}_0 = 0,836 + 0,168 \cdot 9,845 = 2,490 \text{ (тыс. руб.)}$$

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

Найдем доверительный интервал прогноза. Ошибка прогноза

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

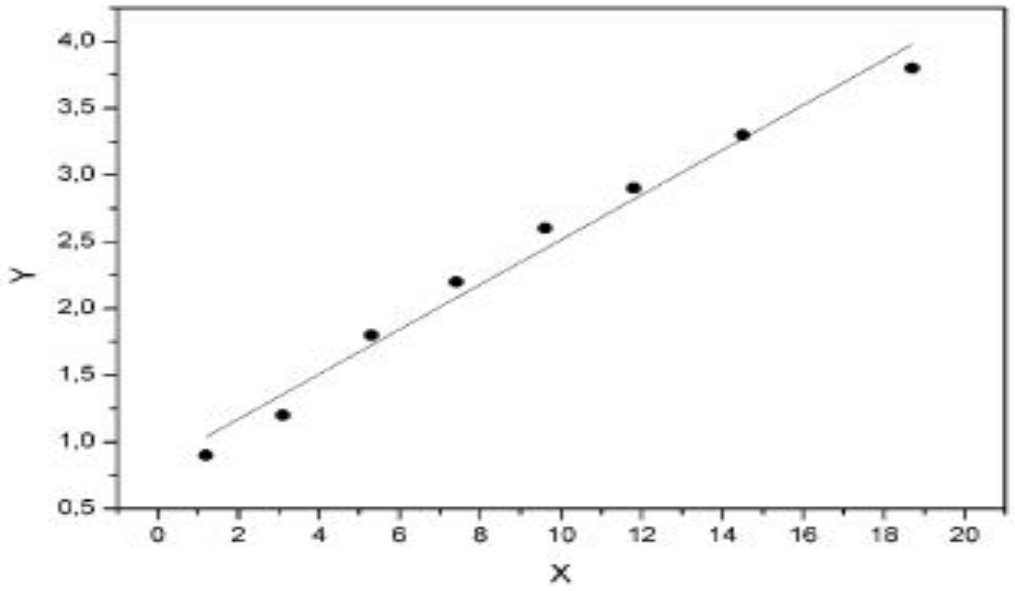
,

а доверительный интервал $(\hat{y}_0 - m_{\hat{y}_0}\cdot t_{\text{табл}} \le y_0^* \le \hat{y}_x + m_{\hat{y}_x}\cdot t_{\text{табл}})$:

$$2,113 < \hat{y}_0^* < 2,867.$$

Т.е. прогноз является статистически надежным.

Теперь на одном графике изобразим исходные данные и линию регрессии:



# Нелинейные модели парной регрессии и корреляции

Различают два класса нелинейных регрессий:

1. Регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, например

– полиномы различных степеней – $\hat{y}_x = a + b \cdot x + c \cdot x^2$,

$\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$;

– равносторонняя гипербола – $\hat{y}_x = a + b/x$;

– полулогарифмическая функция – $\hat{y}_x = a + b \cdot \ln x$.

2. Регрессии, нелинейные по оцениваемым параметрам, например

– степенная – $\hat{y}_x = a \cdot x^b$;

– показательная – $\hat{y}_x = a \cdot b^x$;

– экспоненциальная – $\hat{y}_x = e^{a+b \cdot x}$.

Регрессии нелинейные по оцениваемым параметрам делятся на два типа:

- нелинейные модели внутренне линейные (приводятся к линейному виду с помощью соответствующих преобразований, например, логарифмированием);
- нелинейные модели внутренне нелинейные (к линейному виду не приводятся).

К внутренне линейным моделям относятся, например, степенная функция – $\hat{y}_x = a \cdot x^b$, показательная – $\hat{y}_x = a \cdot b^x$, экспоненциальная – $\hat{y}_x = e^{a+b \cdot x}$, логистическая – $\hat{y}_x = \frac{a}{1+b \cdot e^{-c \cdot x}}$, обратная – $\hat{y}_x = \frac{1}{a+b \cdot x}$.

К внутренне нелинейным моделям можно, например, отнести следующие модели: $\hat{y}_x = a + b \cdot x^c$, $\hat{y}_x = a \cdot \left(1 - \frac{1}{1-x^b}\right)$.

Среди нелинейных моделей наиболее часто используется степенная функция $y = a \cdot x^b \cdot \varepsilon$, которая приводится к линейному виду логарифмированием:

$$\ln y = \ln\left(a \cdot x^b \cdot \varepsilon\right);$$

$$\ln y = \ln a + b \cdot \ln x + \ln \varepsilon;$$

$$Y = A + b \cdot X + \mathrm{E},$$

где $Y = \ln y,\ X = \ln x,\ A = \ln a,\ \mathrm{E} = \ln \varepsilon$. Т.е. МНК мы применяем для преобразованных данных:

$$\begin{cases} A \cdot n + b \cdot \sum X = \sum Y, \\ A \cdot \sum X + b \cdot \sum X^2 = \sum X \cdot Y, \end{cases}$$

а затем потенцированием находим искомое уравнение.

# Формулы для расчета средних коэффициентов эластичности для наиболее часто используемых типов уравнений регрессии

| Вид функции, $y$ | Первая производная, $y'$ | Средний коэффициент эластичности, $\bar{Э}$ |
|---|---|---|
| 1 | 2 | 3 |
| $y = a + b \cdot x + \varepsilon$ | $b$ | $\frac{b \cdot \bar{x}}{a + b \cdot \bar{x}}$ |
| $y = a + b \cdot x + c \cdot x^2 + \varepsilon$ | $b + 2c \cdot x$ | $\frac{(b + 2c \cdot \bar{x}) \cdot \bar{x}}{a + b \cdot \bar{x} + c \cdot \bar{x}^2}$ |
| $y = a + \frac{b}{x} + \varepsilon$ | $-\frac{b}{x^2}$ | $-\frac{b}{a \cdot \bar{x} + b}$ |
| $y = a \cdot x^b \cdot \varepsilon$ | $a \cdot b \cdot x^{b-1}$ | $b$ |
| $y = a \cdot b^x \cdot \varepsilon$ | $a \cdot \ln b \cdot b^x$ | $\bar{x} \cdot \ln b$ |
| $y = a + b \cdot \ln x + \varepsilon$ | $\frac{b}{x}$ | $\frac{b}{a + b \cdot \ln \bar{x}}$ |
| $y = \frac{a}{1 + b \cdot e^{-c \cdot x + e}}$ | $\frac{a \cdot b \cdot c \cdot e^{-c \cdot x}}{(1 + b \cdot e^{-c \cdot x})^2}$ | $\frac{b \cdot c \cdot \bar{x}}{b + e^{c \cdot \bar{x}}}$ |
| $y = \frac{1}{a + b \cdot x + \varepsilon}$ | $-\frac{b}{(a + b \cdot x)^2}$ | $-\frac{b \cdot \bar{x}}{a + b \cdot \bar{x}}$ |

Уравнение нелинейной регрессии, так же, как и в случае линейной зависимости, дополняется показателем тесноты связи. В данном случае это индекс корреляции:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}}, \qquad (1.21)$$

где $\sigma^2_y = \frac{1}{n}\sum (y - \overline{y})^2$ – общая дисперсия результативного признака $y$,

$\sigma^2_{ост} = \frac{1}{n}\sum \left(y - \hat{y}_x\right)^2$ – остаточная дисперсия.

Величина данного показателя находится в пределах: $0 \le \rho_{xy} \le 1$.

Квадрат индекса корреляции носит название индекса детерминации и характеризует долю дисперсии результативного признака $y$, объясняемую регрессией, в общей дисперсии результативного признака:

$$\rho^2_{xy} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_y} = \frac{\sigma^2_{факт}}{\sigma^2_y}, \qquad (1.22)$$

т.е. имеет тот же смысл, что и в линейной регрессии;

$\sigma^2_{факт} = \frac{1}{n}\sum \left(\hat{y}_x - \overline{y}\right)^2$.

Индекс детерминации используется для проверки существенности в целом уравнения регрессии по $F$-критерию Фишера:

$$F = \frac{\rho_{xy}^2}{1-\rho_{xy}^2} \cdot \frac{n-m-1}{m}, \qquad (1.23)$$

где $\rho_{xy}^2$ – индекс детерминации, $n$ – число наблюдений, $m$ – число параметров при переменной $x$. Фактическое значение $F$-критерия (1.23) сравнивается с табличным при уровне значимости $\alpha$ и числе степеней свободы $k_2 = n-m-1$ (для остаточной суммы квадратов) и $k_1 = m$ (для факторной суммы квадратов).