Парная регрессия

1. Понятия регрессионного анализа: зависимые и независимые переменные.
2. Предпосылки применения метода наименьших квадратов (МНК).
3. Свойства оценок метода наименьших квадратов (МНК).
4. Линейная модель парной регрессии. Оценка параметров модели с помощью метода наименьших квадратов (МНК).
5. Показатели качества регрессии модели парной регрессии.
6. Анализ статистической значимости параметров модели парной регрессии.
7. Интервальная оценка параметров модели парной регрессии.
8. Проверка выполнения предпосылок МНК.
9. Интервалы прогноза по линейному уравнению парной регрессии.(Прогнозирование с применением уравнения регрессии).
0. Понятие и причины гетероскедастичности. Последствия гетероскедастичности. Обнаружение гетероскедастичности.
1. Нелинейная регрессия. Нелинейные модели и их линеаризация.

Типы переменных в эконометрической модели

- ***Результирующая (зависимая, эндогенная) переменная Y***

Она характеризует результат или эффективность функционирования экономической системы. Значения ее формируются в процессе и внутри функционирования этой системы под воздействием ряда других переменных и факторов, часть из которых поддается регистрации, управлению и планированию. По своей природе результирующая переменная всегда случайна (стохастична).

- ***Объясняющие (экзогенные, независимые) переменные X***

Это — переменные, которые поддаются регистрации и описывают условия функционирования реальной экономической системы. Они в значительной мере определяют значения результирующих переменных. Еще их называют факторными признаками. В регрессионном анализе это аргументы результирующей функции Y. По своей природе они могут быть как случайными, так и неслучайными.

Регрессионный анализ

Предназначен для исследования зависимости исследуемой переменной от различных факторов и отображения их взаимосвязи в форме регрессионной модели.

- Зависимая (объясняемая) переменная = > ***Y***
- Независимые (объясняющие) переменные =>***X***
- ***По виду функции различают модели:***
  - линейные;
  - нелинейные.
- ***По количеству включенных факторов:***

  - однофакторные (парной регрессии);

  - многофакторные (множественной регрессии).

Предпосылки применения метода наименьших квадратов (МНК)

***Первое условие*. Математическое ожидание случайной составляющей в любом наблюдении должно быть ра** $M(\varepsilon_i) = 0$

Второе условие **состоит в том, что возмущение (или зависимая переменная) есть величина случайная.**

Третье условие **предполагает отсутствие систематической связи между значениями случайной составляющей в любых двух наблюдениях**

$$M(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j)$$

Четвертое условие **означает, что дисперсия случайной составляющей должна быть постоянна для всех наблюдений. Это условие** ***гомоскедастичности*.** $D(\varepsilon_i) = \sigma_\varepsilon^2$

Предположение о нормальности

Наряду с перечисленными условиями Гаусса— Маркова обычно также предполагается нормальность распределения случайного члена.

Свойства оценок метода наименьших квадратов (МНК)

Оценки параметров регрессии должны быть несмещенными, состоятельными и эффективными

Свойства	Интерпретация	Применение
Несмещенность	**Математическое ожидание остатков равно нулю**	**При большом числе выборочных оцениваний остатки не будут накапливаться, оценки можно сравнивать по разным выборкам**
Эффективность	**Оценки считаются эффективными, если они характеризуются наименьшей дисперсией**	**Возможность перехода от точечного оценивания к интервальному**
Состоятельность	**Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки**	**Вероятность получения оценки на заданном расстоянии от истинного значения параметра близка к единице.**

Линейная парная регрессия

$$y_i = a_0 + a_1 \cdot x_i + \varepsilon_i ,$$

где a_0 – постоянная величина,

a_1 – коэффициент регрессии, характеризует угол наклона линии регрессии.

Если $a_1 > 0$, то переменные x и y положительно коррелированы, если $a_1 < 0$ – отрицательно

Или $a_0 + a_1 \cdot x_i$ - неслучайная составляющая;

ε_i – случайная составляющая с нулевым математическим ожиданием и постоянной дисперсией, она учитывает неучтенные факторы, ошибки измерения и пр.

Оценка параметров уравнения регрессии МНК

МНК минимизирует сумму квадратов отклонения фактических значений y_i от расчетных

$$a_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)/(n-1)}{S_x^2} = \frac{\sum(x_i - x)\cdot(y_i - y)}{\sum(x_i - x)^2} =$$

$$= r_{x,y}\cdot\frac{S_y}{S_x} = \frac{\overline{y\cdot x} - \overline{y}\cdot\overline{x}}{\overline{x^2} - \overline{x}^2} = \frac{\sum_{i=1}^{n} y_i\cdot x_i - n\cdot\overline{y}\cdot\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\cdot\overline{x}^2}$$

$$a_0 = \overline{y} - a_1\cdot\overline{x}.$$

$$y_p = a_0 + a_1\cdot x$$

Матричная форма оценки параметров уравнения регрессии МНК

$$Y = X \cdot A + \varepsilon ,$$

где Y – вектор-столбец (n×1) наблюдаемых значений зависимой переменной;

X – матрица (n×2) значений факторов;

A – вектор-столбец (2×1) неизвестных коэффициентов регрессии;

ε – вектор-столбец (n×1) ошибок наблюдений

$$Y = \begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_i \\ \dots & \dots \\ 1 & x_n \end{pmatrix} \quad A = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_i \\ \dots \\ \varepsilon_n \end{pmatrix}$$

- ***Решение системы нормальных уравнений в матричном виде:*** $A = (X' \cdot X)^{-1} \cdot X' \cdot Y$.

Для расчета вектора *A* необходимо:

1. Транспонировать матрицу X => *[ТРАНСП];*
2. Умножить транспонированную матрицу на исходную *(X'X) => [МУМНОЖ];*
3. Вычислить обратную матрицу $(X'X)^{-1}$ => *[МОБР];*

Оценка качества модели регрессии

Качество модели оценивается на основе анализа остаточной компоненты ($\varepsilon_i = y_i - y_p$):

Качество модели регрессии оценивается по следующим направлениям:

- *проверка качества всего уравнения регрессии;*
- *проверка значимости всего уравнения регрессии;*
- *проверка статистической значимости коэффициентов уравнения регрессии;*
- *проверка выполнения предпосылок МНК.*

В основе анализа качества лежит теорема о разложении дисперсии на две составляющие:

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

дисперсия объясненная необъясненная

Разделив обе части уравнения на левую получим:

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

Коэффициент детерминации R^2

Откуда, в окончательном виде имеем :

$$R^2 = \frac{\text{объясняемая сумма квадратов}}{\text{общая сумма квадратов}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\varepsilon_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Коэффициент детерминации показывает долю вариации результативного признака, находящегося под воздействием изучаемых факторов.

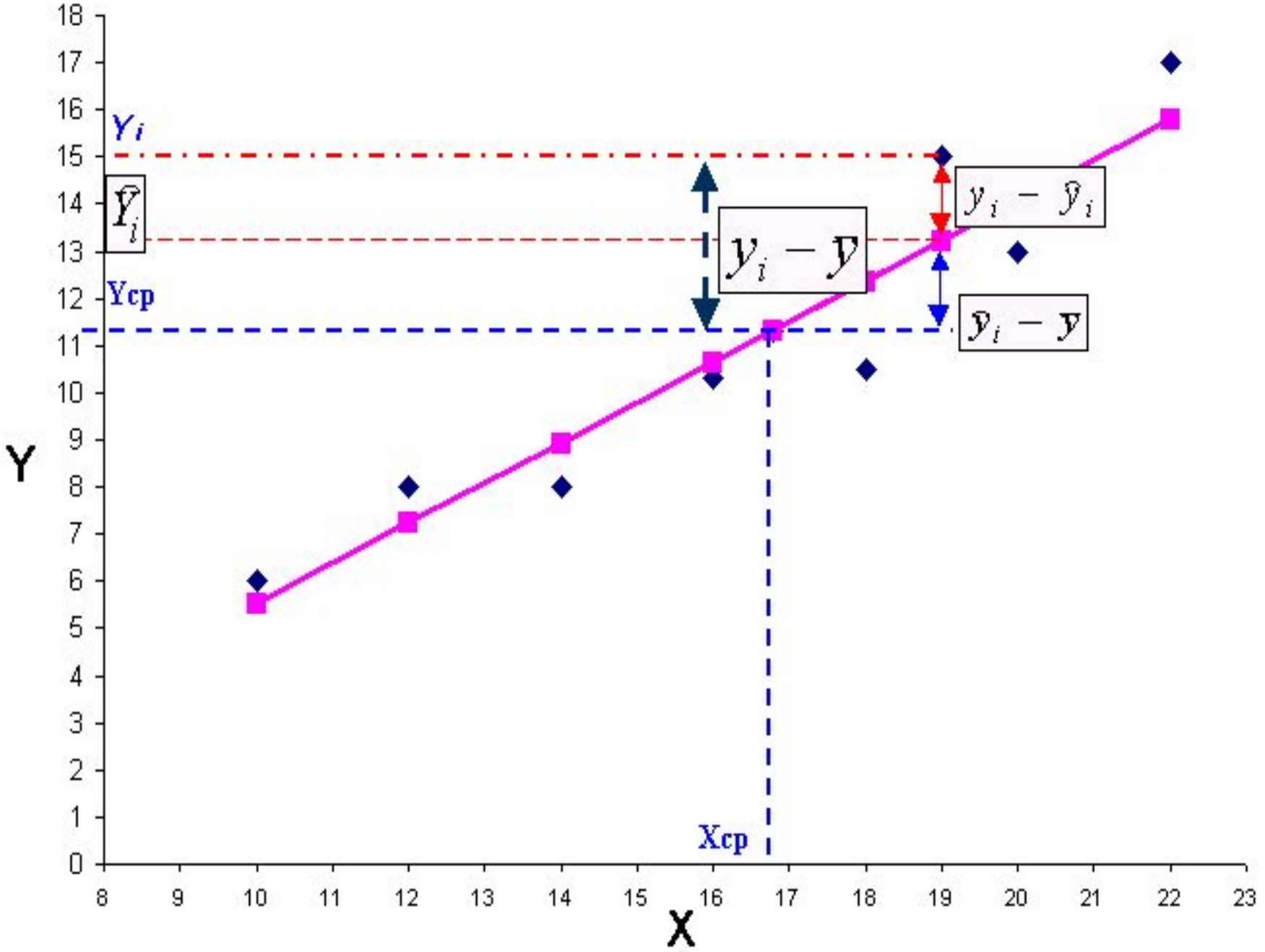
Чем ближе R^2 к *1*, тем выше качество модели.

Если $R^2 = 0$? – связь между признаками отсутствует Если $R^2 = 1$? - связь функциональная

Коэффициент множественной корреляции *R*

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

Он отражает и тесноту связи и точность модели


Yi
$\hat{Y}_i$
Ycp
Xcp
$y_i - \bar{y}$
$y_i - \hat{y}_i$
$\hat{y}_i - \bar{y}$
Y
X
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Для однофакторной модели $R = |r_{y,x}|$.

Критерий Фишера используется для провер-ки значимости модели регрессии при выбранном уровне α и степенях свободы k_1 и k_2.

Для однофакторной модели регрессии:

$$F = \frac{R^2}{1-R^2} \cdot (n-2) = \frac{r_{y,x}^2}{1-r_{y,x}^2} \cdot (n-2)$$

Критерии точности модели

Средняя квадратическая ошибка – (стандартная ошибка оценки)

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2}}$$ - для однофакторной модели

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-k-1}}$$

Если $S_\varepsilon \le \sigma_y$, то модель регрессии использовать целесообразно.

Средняя относительная ошибка аппроксимации:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\varepsilon_i}{y_i}\right| \times 100\%$$

Если $A \le 7\%$, то модель имеет хорошее качество.

Проверка гипотез о значимости параметров уравнения регрессии.

Выдвигается H_0 – гипотеза о незначимом отличии параметра уравнения регрессии от нуля.

Для проверки этой гипотезы используется t – статистика (имеющая распределение Стьюдента).

Расчетные значения ***t*** – критерия определяются по формулам:

$$t_{a0} = |a_0| / S_{a0} \quad \text{и} \quad t_{a1} = |a_1| / S_{a1},$$

где $$S_{a0} = \frac{S_\varepsilon \sqrt{\sum x_i^2}}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}} = S_\varepsilon \frac{\sqrt{\sum x_i^2}}{n \cdot \sigma_x} = \sqrt{\frac{S_\varepsilon^2 \sum x_i^2}{n \cdot \sum (x_i - \bar{x})^2}}$$

$$S_{a1} = \frac{S_\varepsilon \sqrt{n}}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}} = S_\varepsilon \frac{\sqrt{n}}{n \cdot \sigma_x} = \sqrt{\frac{S_\varepsilon^2}{\sum (x_i - \bar{x})^2}}$$

Здесь $$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

t_{a0} или $t_{a1} > t_{табл}$, то параметр значим

[В **Excel** $t_{табл}$ => **СТЬЮДРАСПОБР**]

Интервальная оценка параметров модели

выполняется для значимого уравнения по формулам:

$a_0 = [a_0 \pm t_{табл} \cdot S_{a0}]$ – для свободного члена a_0;

$a_1 = [a_1 \pm t_{табл} \cdot S_{a1}]$ – для параметра a_1.

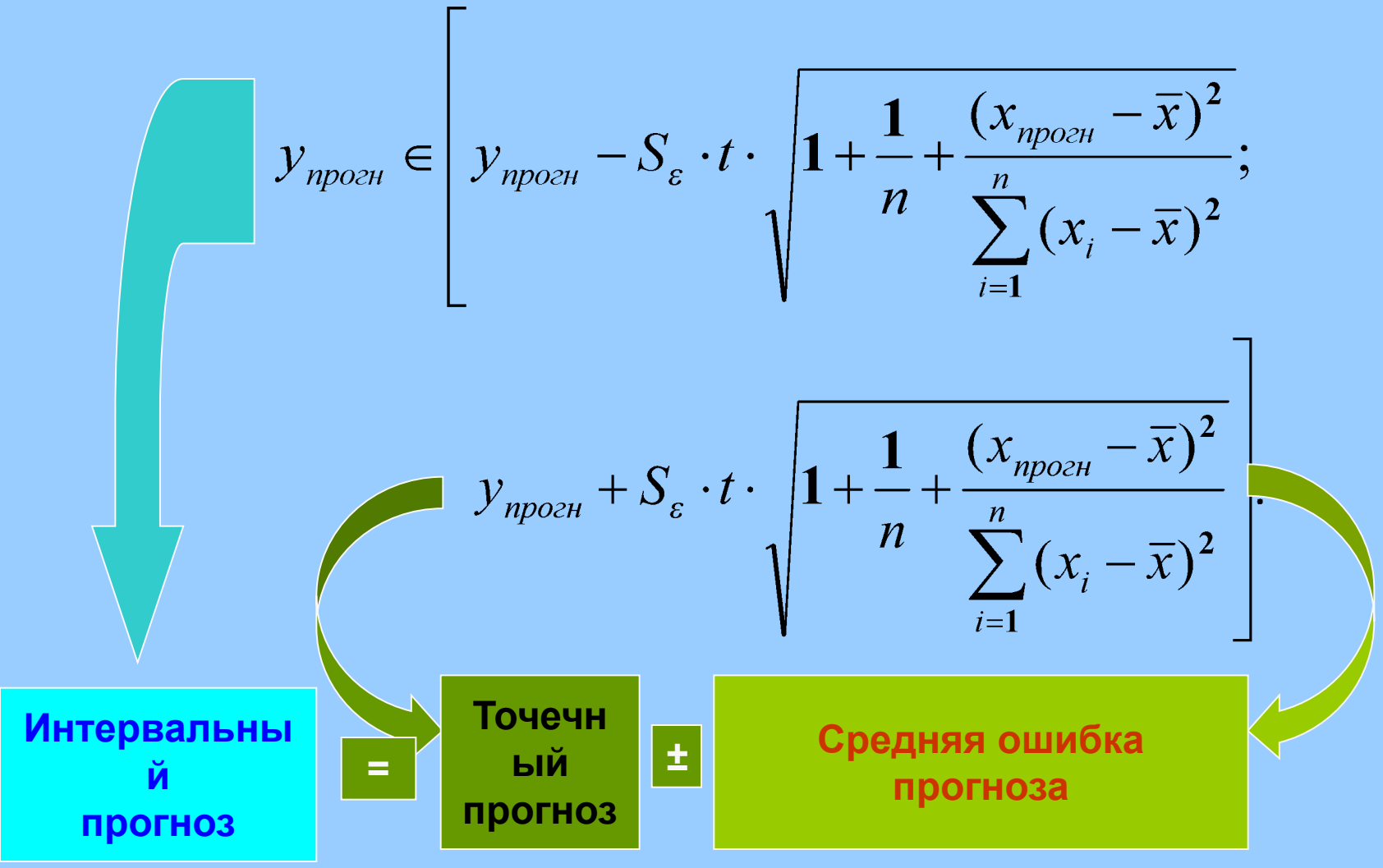
где $t_{табл}$ –критерий Стьюдента для $k = n-2$ степеней,

S_{a0}, S_{a1} – стандартные отклонения

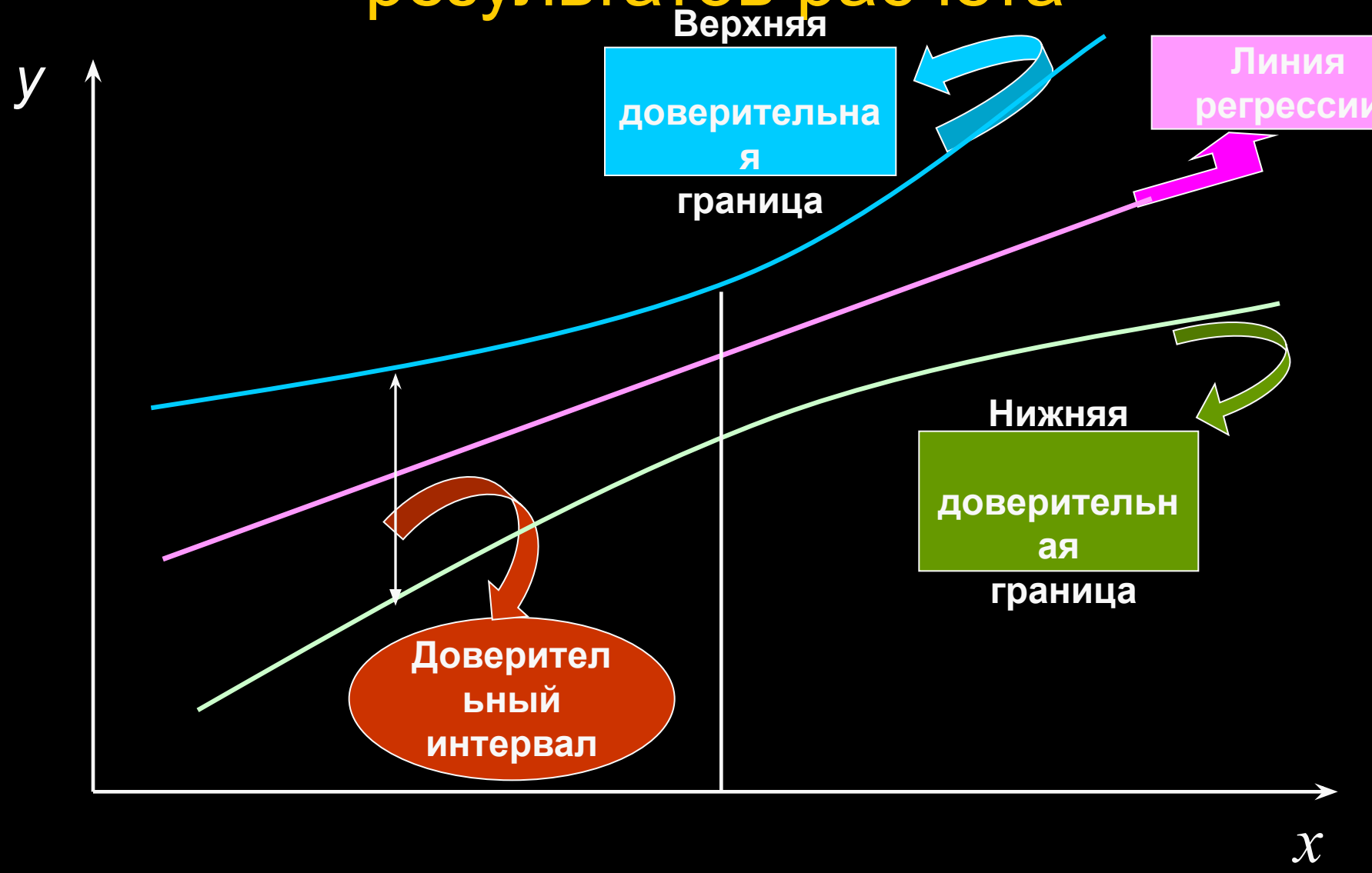
Прогнозирование по уравнению регрессии

Точечный прогноз получают подстановкой ожидаемого значения $x_{прогн}$ в уравнение: $y_{прогн} = a_0 + a_1 \cdot x_{прогн}$

Поскольку вероятность точечного прогноза близка к нулю, то рассчитывается доверительный интервал, в который с вероят-


$y_{прогн} \in \left[y_{прогн} - S_\varepsilon \cdot t \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{прогн} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}; \right.$
$\left. y_{прогн} + S_\varepsilon \cdot t \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{прогн} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right].$
Интервальный прогноз
=
Точечный прогноз
±
Средняя ошибка прогноза

Графическая интерпретация результатов расчета



Регрессионный анализ

- предназначен для исследования зависимости исследуемой переменной от различных факторов и отображения их взаимосвязи в форме регрессионной модели.
- ***В регрессионных моделях*** зависимая переменная *Y* может быть представлена в виде функции *f (X),* где - *X1,X2,…,Xm* независимые (объясняющие) переменные, или факторы.
- Связь между переменной ***Y*** и ***m*** независимыми факторами ***X*** можно охарактеризовать *функцией регрессии* *Y= f (X1,X2,…,Xm)*, которая показывает, каково будет в среднем значение переменной *yi,* если переменные *Xi* примут конкретные значения.

Примеры задач, решаемых с помощью регрессионных моделей

- ***Исследование зависимости заработной платы*** (*Y*) от возраста (*X1*), уровня образования (*X2*), пола (*X3*), стажа работы (*X4*) ($y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$)
- ***Прогноз и планирование выпускаемой продукции по факторам производства*** (производственная функция Кобба – Дугласа означает, что объем выпуска продукции (*Y*), является функцией количества капитала (*K*) и количества (*L*) труда $y = a_0K^{a_1}L^{a_2}$).
- **Прогноз объемов потребления продукции или услуг определенного вида** (кривая Энгеля

$$y = \frac{a_0}{1 + a_1e^{-a_2x}}$$

где *Y* -удельная величина спроса, *X* - среднедушевой доход).

Регрессионные модели с переменной структурой (фиктивные переменные).

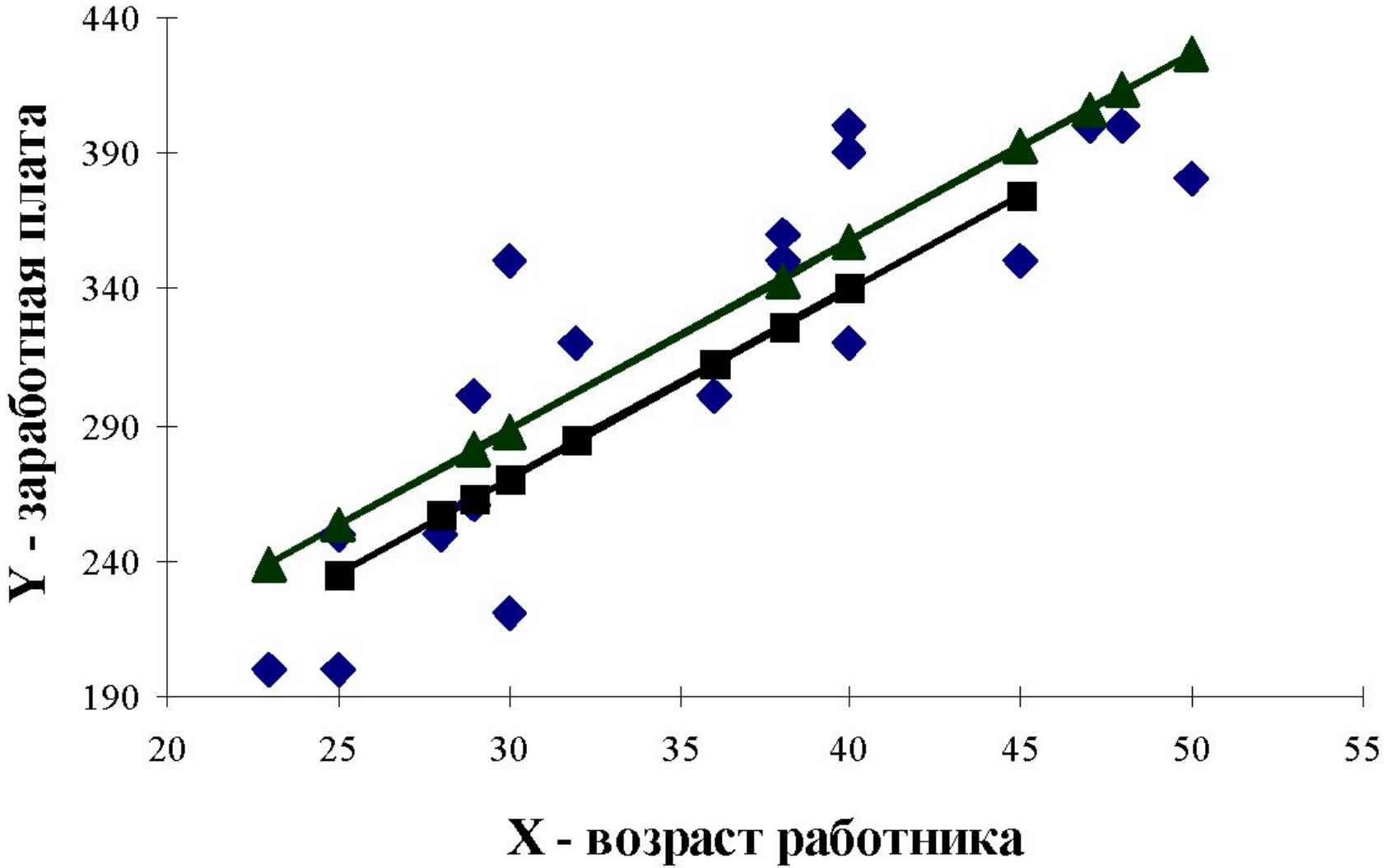
- Построена регрессионная модель зависимости заработной платы работника (Y) от возраста (X) с использованием фиктивной переменной по фактору пол по 20 работникам одного предприятия

$$y = 60,71 + 6,98x + 17,27z$$

- Из полученного уравнения регрессии следует, что при одном и том же возрасте заработная плата у работников мужчин на 17,27$ в месяц выше, чем у женщин.
- Из модели, включающей фиктивную переменную можно получить частные уравнения регрессии для работников мужчин (z=1) и женщин (z=0):

$$y = 77,98 + 6,98x \quad (z = 1)$$

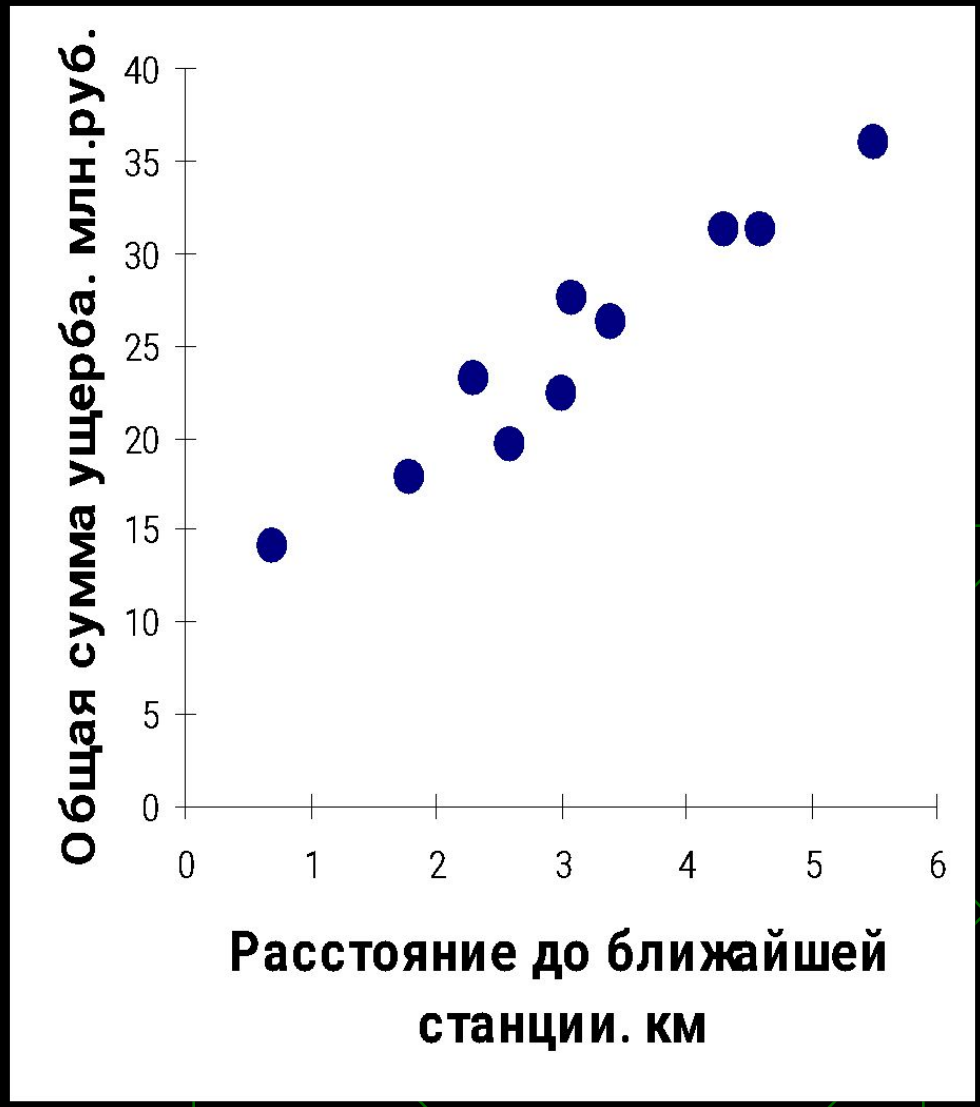
$$y = 60,71 + 6,98x \quad (z = 0).$$


440
390
340
290
240
190
Y - заработная плата
20
25
30
35
40
45
50
55
X - возраст работника

Задача

Администрация страховой компании приняла решение о введении нового вида услуг – страхование на случай пожара. С целью определения тарифов по выборке из 10 случаев пожаров анализируется зависимость стоимости ущерба, нанесенного пожаром от расстояния до ближайшей пожарной станции.

№	Y-Общая сумма ущерба. тыс.руб.	X-Расстояние до ближайшей станции. км
1	26.2	3.4
2	17.8	1.8
3	31.3	4.6
4	23.1	2.3
5	27.5	3.1
6	36	5.5
7	14.1	0.7
8	22.3	3
9	19.6	2.6
10	31.3	4.3



Прогноз по модели Y=10,25+4,69X

Прогноз X

По исходным данным полагают, что расстояние до ближайшей пожарной станции уменьшится на 5% от своего среднего уровня $\overline{x_{км}} = 3.13$

$$x_{км\,прогноз} = 3.13 \cdot 0.95 = 2.97 (\quad)$$

Прогноз Y

$$y_{прогноз} = a + b \cdot x_{прогноз} = 10.25 + 4.69 \cdot 2.97 = 24.2 (\quad . \quad .)$$

тыс. руб

Построение доверительного интервала прогноза

$$U = S_{\hat{y}} \times t_{\alpha} \times \sqrt{1 + 1/n + \frac{(x_{\text{прогноз}} - x_{\text{ср}})^2}{\sum_{i=1}^{n} (x_{i\text{ср}} - x_{\text{ср}})^2}} =$$

$$= 1,801 \times 1,86 \times \sqrt{1 + \frac{1}{10} + \frac{0,026}{17,881}} = 3,51$$

Стандартная ошибка 1.801 $t_{(0,1;\ 8)} = 1,86$

Построение доверительного интервала прогноза

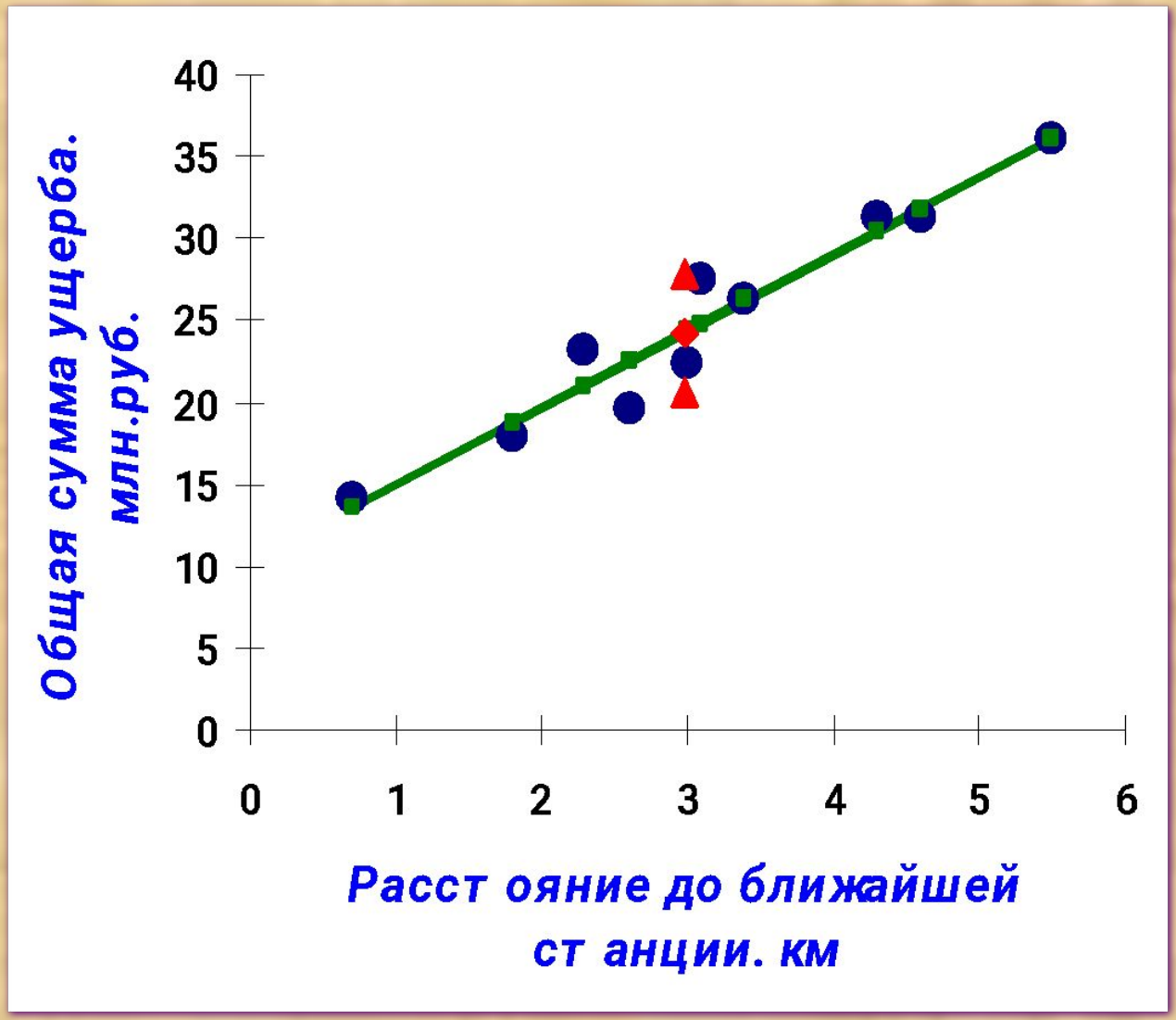
$$U = S_{\hat{y}} \times t_{\alpha} \times \sqrt{1 + 1/n + \frac{(x_{прогноз} - x_{ср})^2}{\sum_{i=1}^{n} (x_{i\text{-}ср} - x)^2}} =$$

$$= 1,801 \times 1,86 \times \sqrt{1 + \frac{1}{10} + \frac{0,026}{17,881}} = 3,51$$

Стандартная ошибка 1.801

Строим доверительный интервал прогноза ущерба с вероятностью 0,90 (t=1,86). Из полученных результатов видно, что интервал от 20,67 до 27,7 тыс. руб. ожидаемой величины ущерба довольно широкий. Значительная неопределенность прогноза линии регрессии, связана, прежде всего с малым объемом выборки (*n=10*), а также тем, что по мере удаления *прогнозного знаения X* от *среднего* ширина доверительного интервала увеличивается.

График прогноза



Задача 1. Задание по эконометрическому моделированию стоимости квартир в Московской области

1. Рассчитайте матрицу парных коэффициентов корреляции; оцените статистическую значимость коэффициентов корреляции.
2. Постройте поле корреляции результативного признака и наиболее тесно связанного с ним фактора.
3. **Рассчитайте параметры линейной парной регрессии.**
4. **Оцените качество каждой модели через коэффициент детерминации, среднюю ошибку аппроксимации и F-критерий Фишера.**
5. **Осуществите прогнозирование среднего значения показателя Y при уровне значимости , если прогнозное значения фактора X составит 80% от его максимального значения. Представьте графически: фактические и модельные значения, точки прогноза.**
6. Используя пошаговую множественную регрессию (метод исключения или метод включения), постройте модель формирования цены квартиры за счёт значимых факторов. Дайте экономическую интерпретацию коэффициентов модели регрессии.
7. Оцените качество построенной модели. Улучшилось ли качество модели по сравнению с однофакторной моделью? Дайте оценку влияния значимых факторов на результат с помощью коэффициентов эластичности, β - и Δ - коэффициентов.

Нелинейная регрессия

При описании экономических процессов могут использоваться также и нелинейные функции.

Различают два класса нелинейных регрессий:

- Нелинейные относительно объясняющих переменных, но линейные по оцениваемым параметрам:
  - Полиномы разных степеней

$$y_i = a_0 + a_1 \cdot x_i + a_2 \cdot x_i^2 + a_3 \cdot x_i^3 + \ldots + a_k \cdot x_i^k + \varepsilon_i$$

  - Равносторонняя гипербола $y_i = a_0 + a_1 / x_i + \varepsilon_i$.
- Нелинейные по оцениваемым параметрам:
  - Степенная $y_i = a_0 \cdot x_i^{a1} \cdot \varepsilon_i$

кривые спроса,предложения, Энгеля, производственные функции,

кривые освоения, зависимость вал. Нац. Прод. От уровня занятости

- Показательная $y_i = a_0 \cdot a_1^{xi} \cdot \varepsilon_i$
- Экспоненциальная $y_i = e^{a0 + a1 \cdot xi} \cdot \varepsilon_i$

- Первый класс нелинейных моделей легко сводится к линейным путем замены нелинейных переменных x^k новыми линейными переменными z_k и затем применяют МНК.
- Во втором классе выделяют два подкласса:
  - Внутренне линейные – путем преобразований сводятся к линейному виду;
  - Внутренне нелинейные – путем логарифмирования приводятся к линейному виду, либо используются итеративные процедуры оценки параметров.

Остальное см. практику