Федеральное государственное автономное
образовательное учреждение высшего образования
«Научно-образовательная корпорация ИТМО»

ФАКУЛЬТЕТ ПРОГРАММНОЙ ИНЖЕНЕРИИ И КОМПЬЮТЕРНОЙ ТЕХНИКИ

Индивидуальное домашнее задание №7 «Построение оценки линейной регрессии»

Вариант № 2 (77)

Работу выполнили:
студент группы P3209
Зайцева И. С.
студент группы P3217
Русакова Е. Д.

Преподаватель:
Милованович Е. В.

г. Санкт-Петербург
2024 г.

Цель работы:

На основании анализа выборки:

1. Предположить вид модели
2. Точечно оценить параметры данной модели
3. Проверить адекватность модели
4. Интервально оценить функцию регрессии и ее параметров

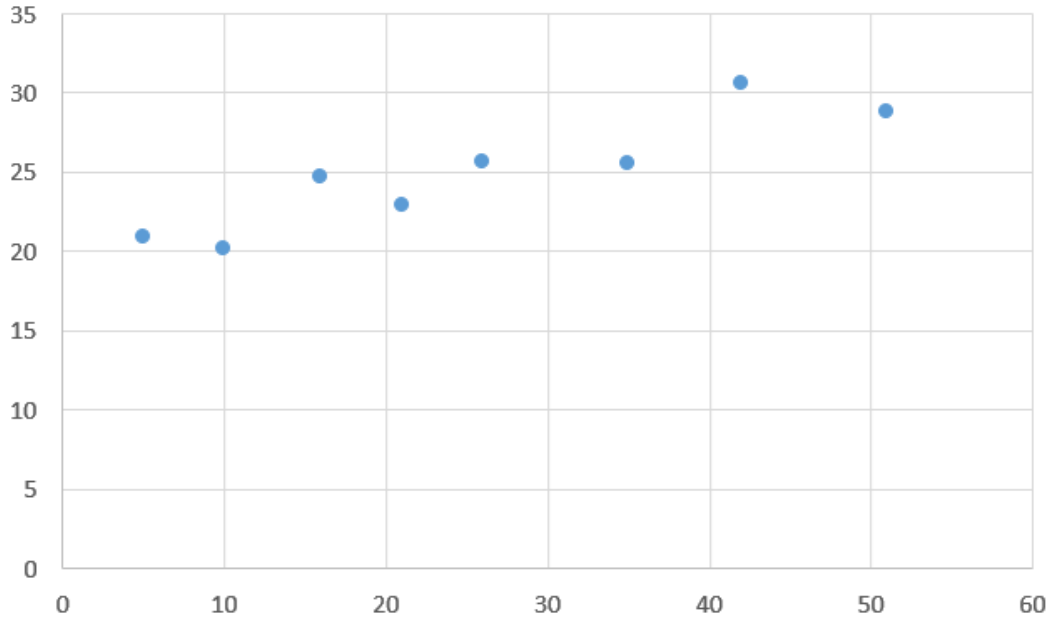
Исходные данные:

x	5	10	16	21	26	35	42	51
y	20,9	20,1	24,7	22,9	25,6	25,5	30,6	28,8

$n = 8$

Ход работы:

Представим на графике исходные данные:



Выберем функцию вида

$$y = a_0 + a_1 x$$

Найдем точечные оценки неизвестных параметров a_0, a_1 функции регрессии методом средних:
Два неизвестных параметра, таблицу данных делим на две части:

$$+\begin{cases} 20.9 = \widetilde{a_0} + 5\widetilde{a_1} \\ 20.1 = \widetilde{a_0} + 10\widetilde{a_1} \\ 24.7 = \widetilde{a_0} + 16\widetilde{a_1} \\ 22.9 = \widetilde{a_0} + 21\widetilde{a_1} \end{cases}$$

$$88.6 = 4\widetilde{a_0} + 52\widetilde{a_1}$$

$$+\begin{cases} 25.6 = \widetilde{a_0} + 26\widetilde{a_1} \\ 25.5 = \widetilde{a_0} + 35\widetilde{a_1} \\ 30.6 = \widetilde{a_0} + 42\widetilde{a_1} \\ 28.8 = \widetilde{a_0} + 51\widetilde{a_1} \end{cases}$$

$$110.5 = 4\widetilde{a_0} + 154a_1$$

$$\begin{cases} 88.6 = 4\widetilde{a_0} + 52\widetilde{a_1} \\ 110.5 = 4\widetilde{a_0} + 154\widetilde{a_1} \end{cases}$$

$$\begin{cases} \widetilde{a_0} = 19.359 \\ \widetilde{a_1} = 0.215 \end{cases}$$

Найдем точечные оценки неизвестных параметров a_0, a_1 функции регрессии методом наименьших квадратов:

$$\tilde{y} = \widetilde{a_0} + \widetilde{a_1}x$$

$$S = \sum (y_i - \widetilde{a_0} - \widetilde{a_1}x_i)^2$$

Минимизируем критерий близости S оценки функции регрессии к экспериментальным данным:

$$\begin{cases} \dfrac{\partial S}{\partial \widetilde{a_0}} = 2\sum (y_i - \widetilde{a_0} - \widetilde{a_1}x_i)(-1) = 0 \\ \dfrac{\partial S}{\partial \widetilde{a_1}} = 2\sum (y_i - \widetilde{a_0} - \widetilde{a_1}x_i)(-x_i) = 0 \end{cases}$$

$$\begin{cases} \sum (-1)y_i + \sum \widetilde{a_0} + \sum \widetilde{a_1}x_i = 0 \\ \sum (-x_i)y_i + \sum \widetilde{a_0}\, x_i + \sum \widetilde{a_1}x_i^2 = 0 \end{cases}$$

$$\begin{cases} 8\widetilde{a_0} + \widetilde{a_1}\sum x_i = \sum y_i \\ \widetilde{a_0}\sum x_i + \widetilde{a_1}\sum x_i^2 = \sum x_i y_i \end{cases}$$

$$\begin{cases} 8\widetilde{a_0} + 206\widetilde{a_1} = 199{,}1 \\ 206\widetilde{a_0} + 7088\widetilde{a_1} = 5493{,}7 \end{cases}$$

$$\begin{cases} \widetilde{a_0} = 19{,}59 \\ \widetilde{a_1} = 0{,}2057 \end{cases}$$

$$S_{\min}^{(1)} = \sum (y_i - \widetilde{a_0} - \widetilde{a_1}x_i)^2 = \sum (y_i - 19{,}59 - 0{,}2057x_i)^2 = 16{,}1607$$

Проверка статистической гипотезы об адекватности модели экспериментальных данных

$$S_{\min}^{(1)} = \sum (y_i - \widetilde{a_0} - \widetilde{a_1}x_i)^2$$

$$S_{\min}^{(2)} = \sum (y_i - \widetilde{a_0} - \widetilde{a_1}x_i - \widetilde{a_2}x_i^2)^2$$

1. H_0 – модель (1) может считаться адекватной
   H_1 – модель (1) не адекватная
2. Уровень значимости α = 0,05
3. Критерий Фишера

Найдем точечные оценки неизвестных параметров для модели (2):

$$S = \sum \left(y_i - \widetilde{a_0} - \widetilde{a_1}x_i - \widetilde{a_2}x_i^2\right)^2$$

$$\begin{cases} \frac{\partial S}{\partial \widetilde{a_0}} = 2\sum(y_i - \widetilde{a_0} - \widetilde{a_1}x_i - \widetilde{a_2}x_i^2)(-1) = 0 \\ \frac{\partial S}{\partial \widetilde{a_1}} = 2\sum(y_i - \widetilde{a_0} - \widetilde{a_1}x_i - \widetilde{a_2}x_i^2)(-x_i) = 0 \\ \frac{\partial S}{\partial \widetilde{a_2}} = 2\sum(y_i - \widetilde{a_0} - \widetilde{a_1}x_i - \widetilde{a_2}x_i^2)(-x_i^2) = 0 \end{cases}$$

$$\begin{cases} \sum(-1)y_i + \sum \widetilde{a_0} + \sum \widetilde{a_1}x_i + \sum \widetilde{a_2}x_i^2 = 0 \\ \sum(-x_i)y_i + \sum \widetilde{a_0}\,x_i + \sum \widetilde{a_1}x_i^2 + \sum \widetilde{a_2}x_i^3 = 0 \\ \sum(-x_i^2)y_i + \sum \widetilde{a_0}\,x_i^2 + \sum \widetilde{a_1}x_i^3 + \sum \widetilde{a_2}x_i^4 = 0 \end{cases}$$

$$\begin{cases} 8\widetilde{a_0} + \widetilde{a_1}\sum x_i + \widetilde{a_2}\sum x_i^2 = \sum y_i \\ \widetilde{a_0}\sum x_i + \widetilde{a_1}\sum x_i^2 + \widetilde{a_2}\sum x_i^3 = \sum x_i y_i \\ \widetilde{a_0}\sum x_i^2 + \widetilde{a_1}\sum x_i^3 + \widetilde{a_2}\sum x_i^4 = \sum x_i^2 y_i \end{cases}$$

$$\begin{cases} 8\widetilde{a_0} + 206\widetilde{a_1} + 7088\widetilde{a_2} = 199{,}1 \\ 206\widetilde{a_0} + 7088\widetilde{a_1} + 281672\widetilde{a_2} = 5493{,}7 \\ 7088\widetilde{a_0} + 281672\widetilde{a_1} + 12105140\widetilde{a_2} = 196385 \end{cases}$$

$$\begin{cases} \widetilde{a_0} = 18{,}867 \\ \widetilde{a_1} = 0{,}279 \\ \widetilde{a_2} = -0{,}001 \end{cases}$$

$$S_{\min}^{(2)} = \sum(y_i - \widetilde{a_0} - \widetilde{a_1}x_i - \widetilde{a_2}x_i^2)^2 = S_{\min}^{(2)} = \sum(y_i - 18{,}867 - 0{,}279x_i + 0{,}001x_i^2)^2 = 15{,}6106$$

Статистический критерий:

$$F = \frac{\frac{1}{k-m} * \left(S_{min}^{(1)} - S_{min}^{(2)}\right)}{\frac{1}{n-k-1} * S_{min}^{(2)}}, \text{ где } n = 8, k = 3, m = 2$$

$$F = \frac{\frac{1}{k-m} * \left(S_{min}^{(1)} - S_{min}^{(2)}\right)}{\frac{1}{n-k-1} * S_{min}^{(2)}} = \frac{\frac{1}{3-2} * (16{,}1607 - 15{,}6106)}{\frac{1}{8-3-1} * 15{,}6106} = \frac{4 * (0{,}550099)}{15{,}6106} = 0{,}14$$

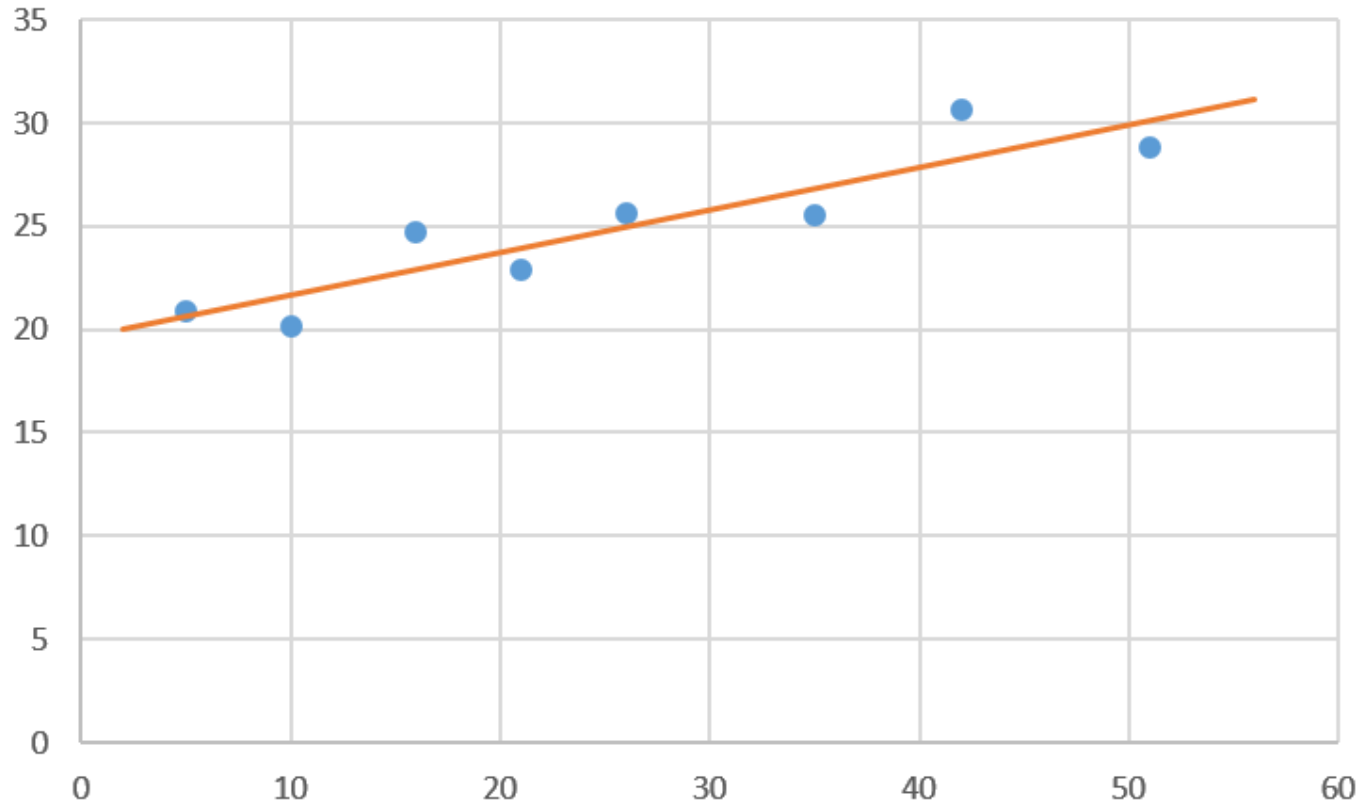
F имеет распределение Фишера с числом степеней свободы $k - m = 1$, $n - k - 1 = 4$

$$f_{\text{крит}} = f_{0{,}05;1;4} = 7{,}709$$

Допустимая область $(0; f_{\text{крит}})$
Критическая область $(f_{\text{крит}}; +\infty)$

$F \in (0; f_{\text{крит}})$, следовательно гипотезу H_0 следует принять.

Функция регрессии имеет вид линейной регрессии $y = a_0 + a_1 x$



Построим доверительные интервалы для коэффициентов функции регрессии:
Доверительная вероятность $\beta = 0{,}9$
Ошибки измерения y_i имеют нормальное распределение с математическим ожиданием $M(\varepsilon_i) = 0$, а $D(\varepsilon_i) = const$, но неизвестна.
Можно утверждать, что $y_i = a_0 + a_1 x_i + \varepsilon_i$.
Случайные величины a_0 u a_1 линейно зависят от величины ε_i => они тоже имеют нормальное распределение.

$$\tilde{\sigma}^2[\varepsilon_i] = \tilde{\sigma}^2$$

$$\tilde{\sigma}^2 = \frac{S_{min}}{n-2} = \frac{16.1607}{8-2} = 2.693$$

$\widetilde{D}(\tilde{a}_i)$ может быть найдена через матрицу оценок корреляционных моментов:

$$K = \begin{pmatrix} K(a_0, a_0) = D(a_0) & K(a_0, a_1) \\ K(a_1, a_0) & K(a_1, a_1) = D(a_1) \end{pmatrix}$$

Воспользуемся утверждением из лекции: $\widetilde{K} = \tilde{\sigma}^2 P^{-1}$

$$P = \begin{pmatrix} n & \sum_{i=0}^{n} x_i \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \end{pmatrix} = \begin{pmatrix} 8 & 206 \\ 206 & 7088 \end{pmatrix}$$

$$P^{-1} = \begin{pmatrix} \frac{1772}{3567} & -\frac{103}{7134} \\ -\frac{103}{7134} & \frac{2}{3567} \end{pmatrix} = \begin{pmatrix} 0.49678 & -0.0144 \\ -0.0144 & 0.00056 \end{pmatrix}$$

$$\widetilde{K} = \tilde{\sigma}^2 P^{-1} = 2.69345 * \begin{pmatrix} 0.49678 & -0.0144 \\ -0.0144 & 0.00056 \end{pmatrix} = \begin{pmatrix} 1.338 & -0.0399 \\ -0.039 & 0.0015 \end{pmatrix}$$

$$=> \widetilde{D}(\widetilde{a_0}) = 1.338,\ \widetilde{D}(\widetilde{a_1}) = 0.0015$$

$$\tilde{\sigma}(\widetilde{a_0}) = \sqrt{\widetilde{D}(\widetilde{a_0})} = 1.157$$

$$\tilde{\sigma}(\widetilde{a_1}) = \sqrt{\widetilde{D}(\widetilde{a_1})} = 0.039$$

Доверительные интервалы: $\widetilde{a_i} - t\tilde{\sigma}(\widetilde{a_i}) < a_i < \widetilde{a_i} + t\tilde{\sigma}(\widetilde{a_i})$

$$\widetilde{a_0} - t\tilde{\sigma}(\widetilde{a_0}) < a_0 < \widetilde{a_0} + t\tilde{\sigma}(\widetilde{a_0})$$
$$\widetilde{a_1} - t\tilde{\sigma}(\widetilde{a_1}) < a_1 < \widetilde{a_1} + t\tilde{\sigma}(\widetilde{a_1})$$

По таблице распределения Стьюдента:

$$t_{(0.9,6)} = 1.44$$

$$19.59 - 1.44 * 1.157 < a_0 < 19.59 + 1.44 * 1.157$$
$$0.2057 - 1.44 * 0.039 < a_1 < 0.2057 + 1.44 * 0.039$$

$$17.925 < a_0 < 21.256$$
$$0.206 < a_1 < 0.262$$

Доверительный интервал для всей функции:

$$t_{(0.9,6)} = 1.44$$
$$\tilde{y}(x) - t\tilde{\sigma}\big(\tilde{y}(x)\big) < M(Y/X = x) < \tilde{y}(x) + t\tilde{\sigma}\big(\tilde{y}(x)\big)$$

Найдем $\tilde{\sigma}\big(\tilde{y}(x)\big)$:

$$\widetilde{K}(\widetilde{a_0}, \widetilde{a_1}) = -0.0389$$

$$\tilde{\sigma}\big(\tilde{y}(x)\big) = \tilde{\sigma}(\widetilde{a_0} + \widetilde{a_1}x) = \sqrt{\widetilde{D}(\widetilde{a_0}) + 2\widetilde{K}(\widetilde{a_0}, \widetilde{a_1})x + \widetilde{D}(\widetilde{a_1})x^2}$$
$$= \sqrt{1.338 + 2 * (-0.0389) * x + 0.0015 * x^2} = \sqrt{1.338 - 0.0778 * x + 0.0015 * x^2}$$

$x = 5$:

$$\tilde{\sigma}\big(\tilde{y}(5)\big) = \sqrt{1.338 - 0.0778 * 5 + 0.0015 * 5^2} = 0.993$$
$$\tilde{y}(5) = \widetilde{a_0} + \widetilde{a_1}\, x = \widetilde{a_0} + \widetilde{a_1} * 5 = 20{,}619$$

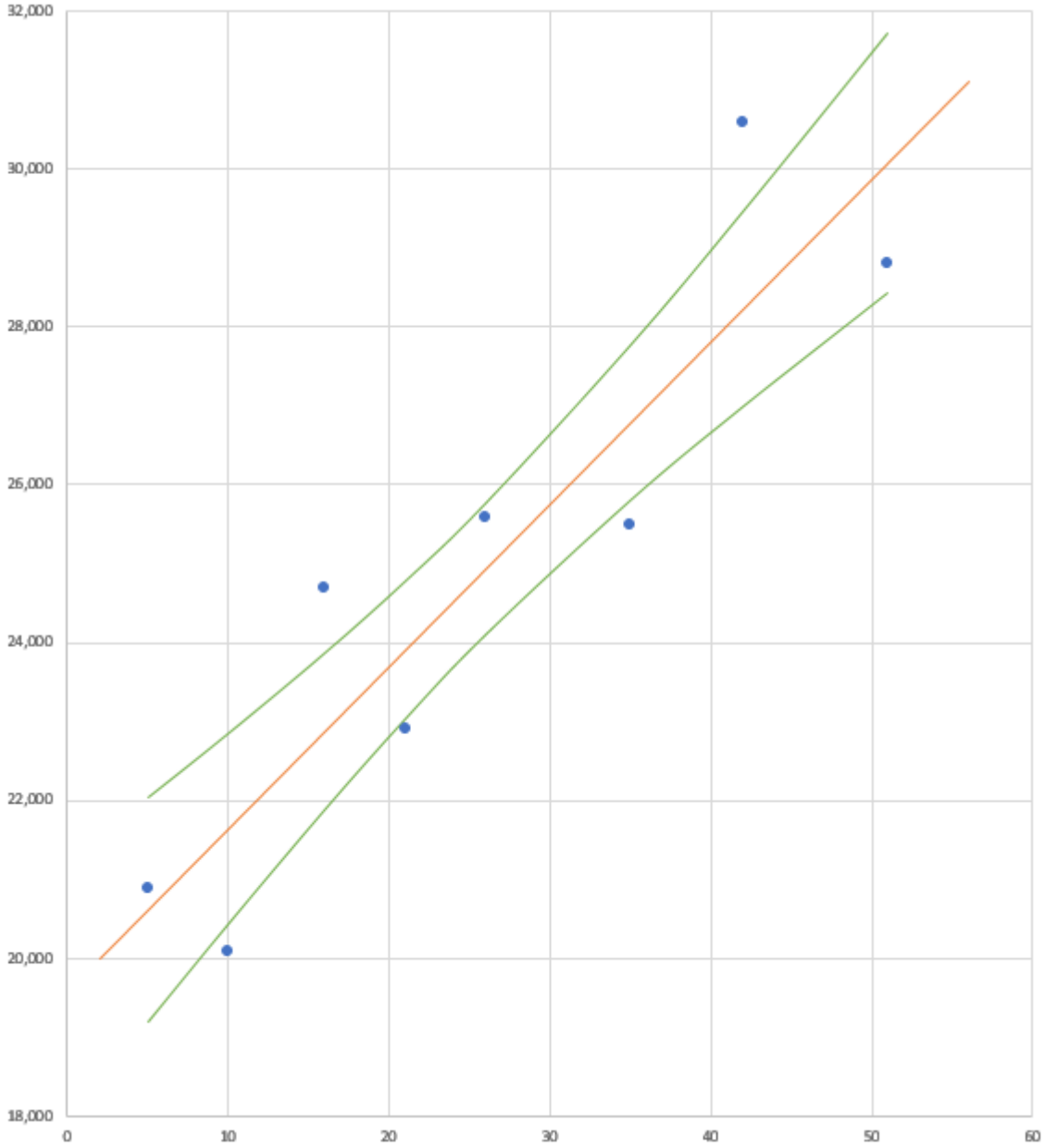
Аналогично заполним следующую таблицу:

x	$\tilde{y}(x)$	$\tilde{\sigma}\big(\tilde{y}(x)\big)$
5	20,619	0,993
10	21,648	0,843
16	22,882	0,693
21	23,910	0,609
26	24,939	0,580
35	26,790	0,683
42	28,230	0,858
51	30,082	1,140

Получим доверительные интервалы для всей функции при различных значениях x:

$\tilde{y}(x) - t\tilde{\sigma}(\tilde{y}(x)) < M(Y/X = x) < \tilde{y}(x) + t\tilde{\sigma}(\tilde{y}(x))$		
19,189	M[Y/X=5]	22,050
20,433	M[Y/X=10]	22,862
21,884	M[Y/X=16]	23,880
23,034	M[Y/X=21]	24,787
24,103	M[Y/X=26]	25,775
25,807	M[Y/X=35]	27,773
26,995	M[Y/X=42]	29,465
28,440	M[Y/X=51]	31,723

Получи итоговый график, где изображены исходные данные (синий), линия регрессии (оранжевый), доверительный интервал (зеленый):



Результаты:

Линия регрессии имеет вид: $y = a_0 + a_1 x$

Доверительные интервалы для коэффициентов функции регрессии:

$$17.925 < a_0 < 21.256$$

$$0.206 < a_1 < 0.262$$

Доверительные интервалы для всей функции:

$\tilde{y}(x) - t\tilde{\sigma}(\tilde{y}(x)) < M(Y/X = x) < \tilde{y}(x) + t\tilde{\sigma}(\tilde{y}(x))$		
19,189	M[Y/X=5]	22,050
20,433	M[Y/X=10]	22,862
21,884	M[Y/X=16]	23,880
23,034	M[Y/X=21]	24,787
24,103	M[Y/X=26]	25,775
25,807	M[Y/X=35]	27,773
26,995	M[Y/X=42]	29,465
28,440	M[Y/X=51]	31,723

Вывод:

Проанализировав исходные данные, мы выяснили вид линии регрессии, нашли доверительные интервалы для функции регрессии и ее коэффициентов.