УНИВЕРСИТЕТ ИННОПОЛИС

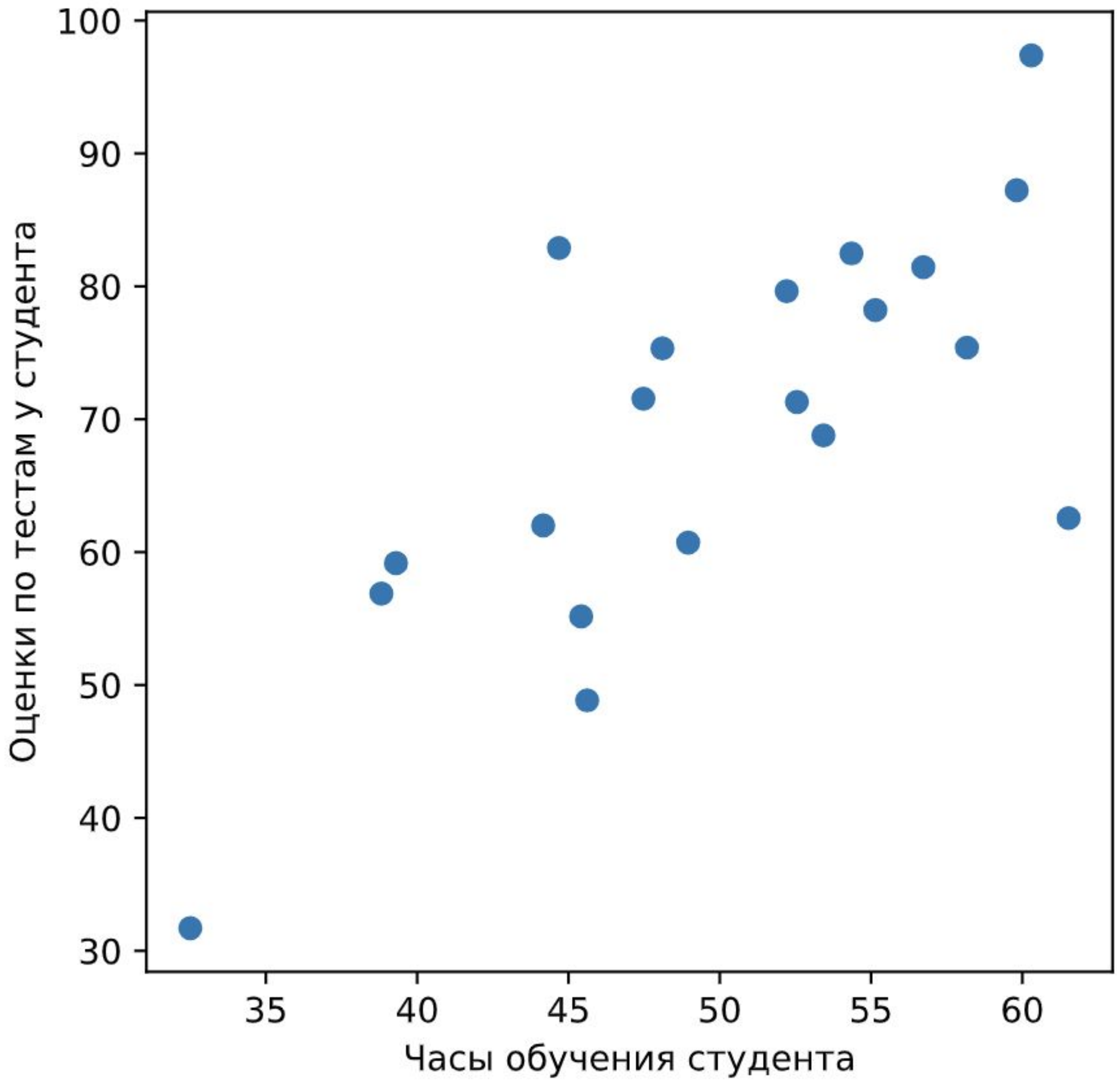
Линейная регрессия

Часть 1

Воробьёва Мария

- maria.vorobyova.ser@gmail.com
- @SparrowMaria

Парная линейная регрессия



У нас есть данные, мы построили график и увидели следующее…

Что делать?

Кажется, что в данных прослеживается закономерность…

И она похоже на прямую…

мы знаем, что уравнение прямой на плоскости
y = a + b * x

Парная линейная регрессия

Уравнение прямой - это и есть линейная регрессия, в данном случае парная линейная регрессия

$$Y = a + b \cdot x$$

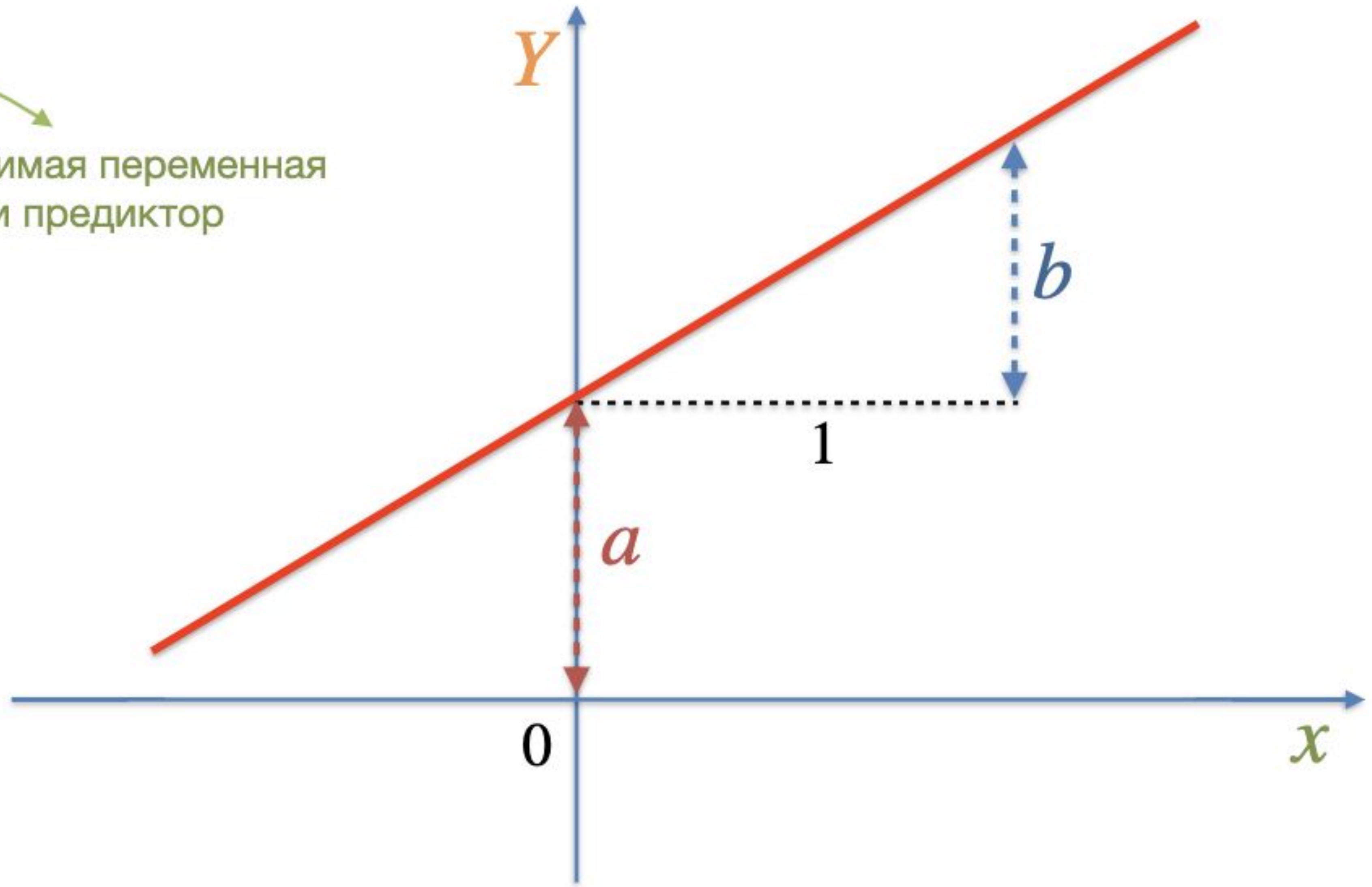
зависимая переменная или переменная отклика

свободный член линии оценки

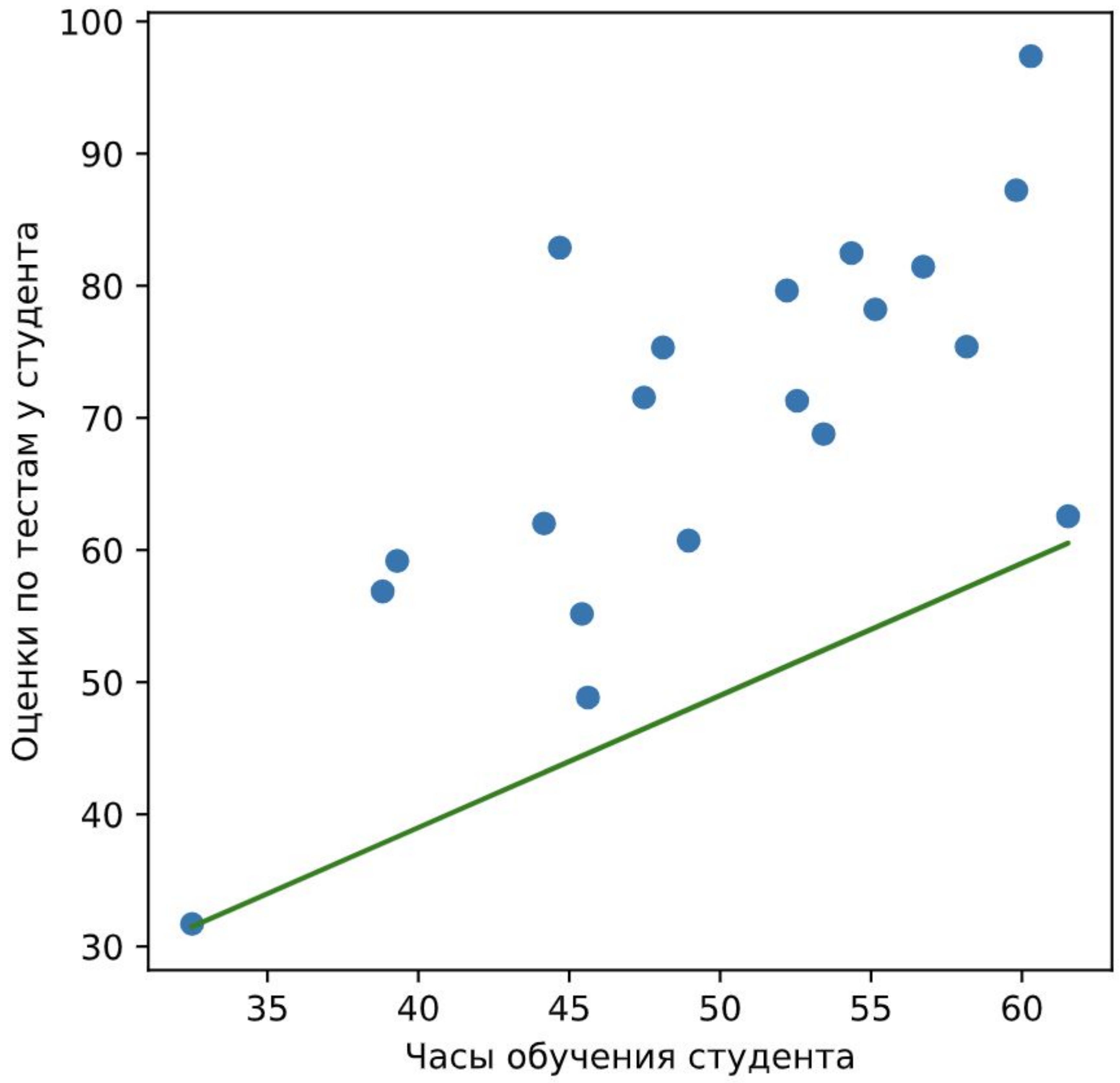
угловой коэффициент

независимая переменная или предиктор

коэффициенты регрессии оценённой линии



Парная линейная регрессия

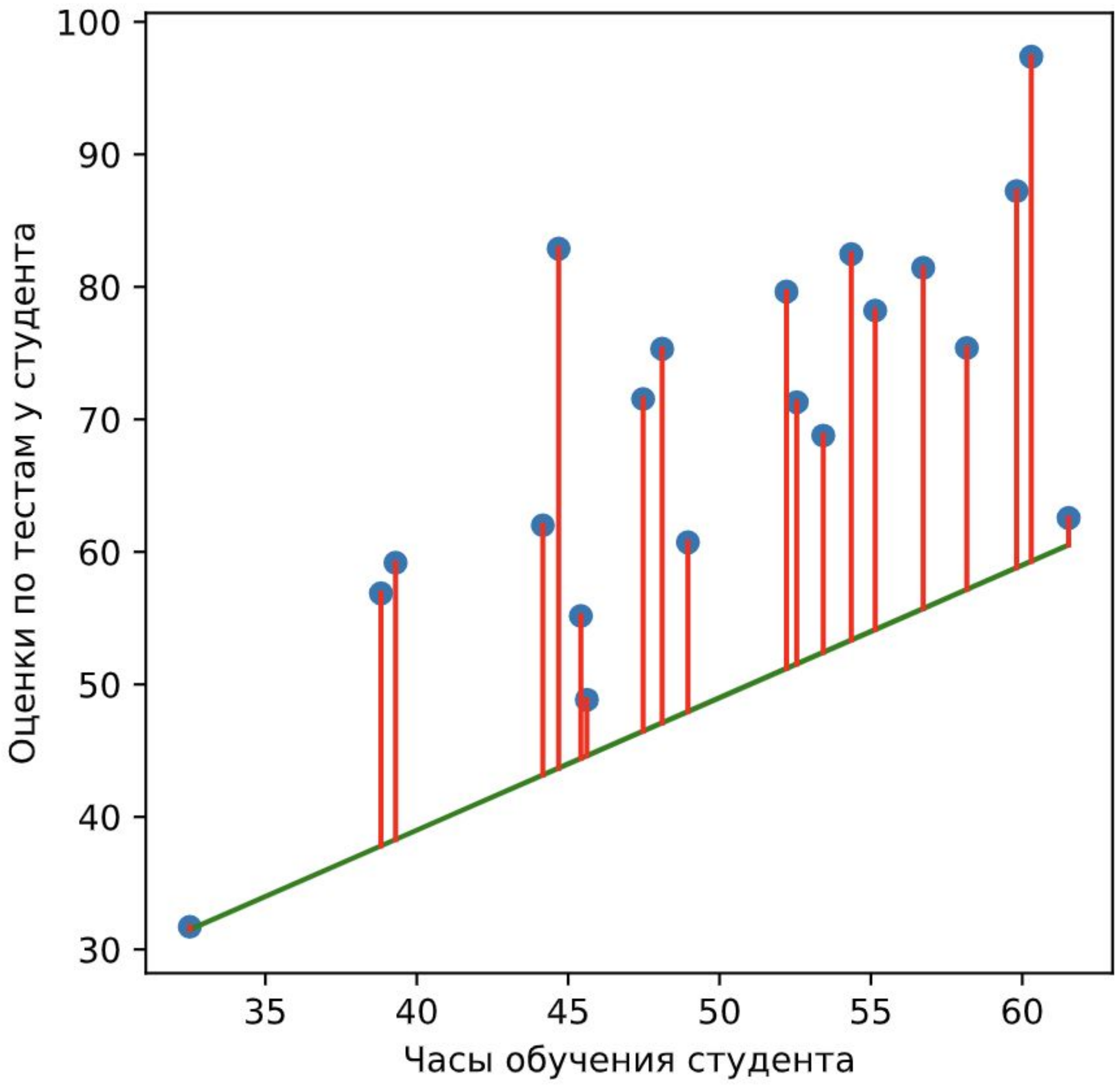


Отлично, построили прямую…

Но кажется, что-то не совсем то, что нам хотелось бы

А что нам не нравится?

Парная линейная регрессия



Попробуем оценить, на сколько наша прямая “не попадает” в наши точки

Посчитаем разности между фактическими данными и точками на прямой,

то есть посчитаем абсолютные суммы длин красных отрезков, получим 20.47

Кажется, можно лучше

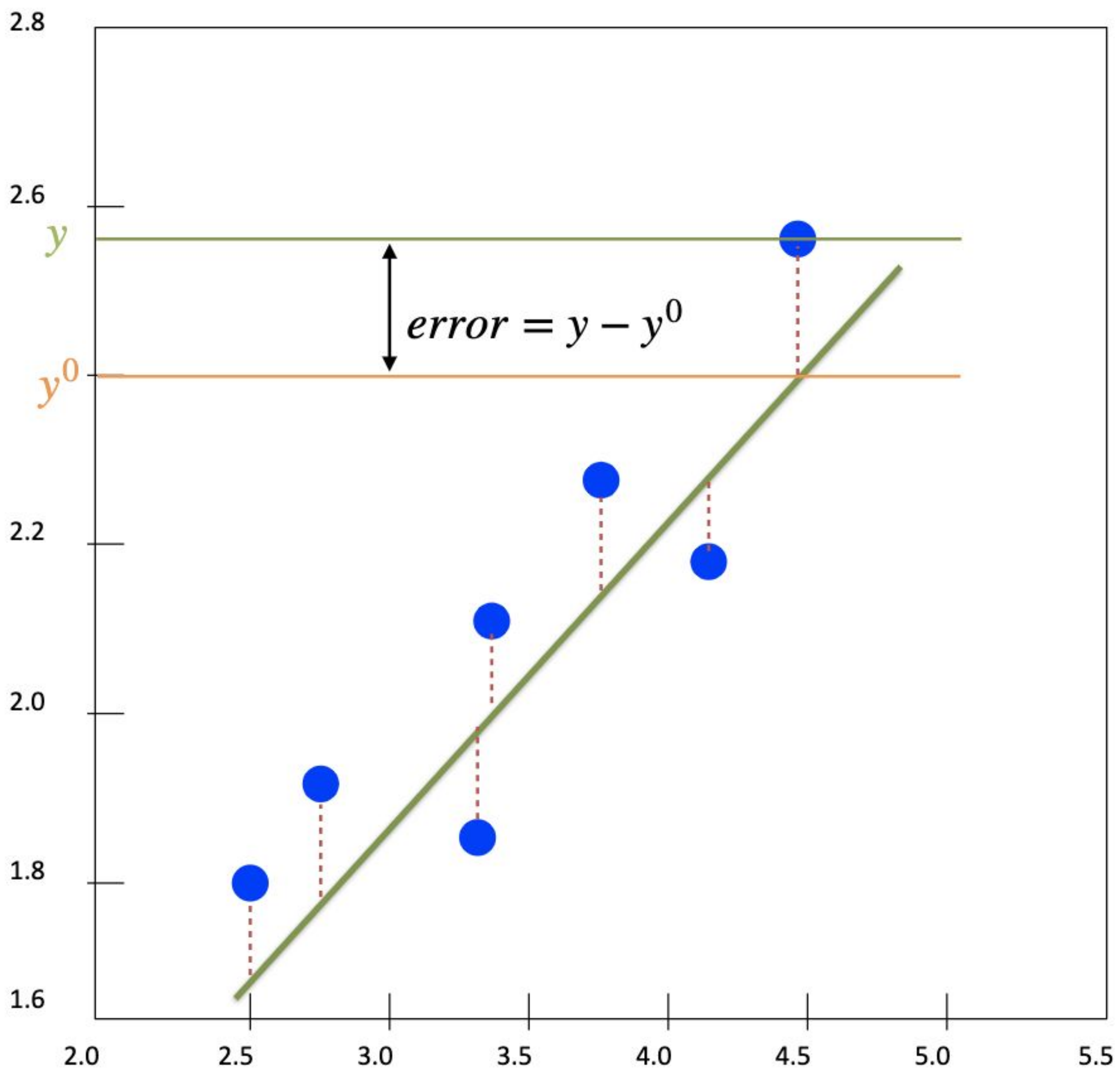
Парная линейная регрессия

Мы должны построить такую прямую, для которой отклонения от фактических точек будут минимальны

$$\sum_{i=1}^{N}(y_i - y_i^0)^2 =$$

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 =$$

$$\sum_{i=1}^{N}(y_i - a - bx_i)^2$$



Парная линейная регрессия

Необходимо минимизировать сумму квадратов отклонений RSS (Residiual Sum of Squares)

$$\text{RSS} = \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

Для того, чтобы определить прямую, необходимо найти a и b

нам поможет Метод Наименьших Квадратов МНК

Парная линейная регрессия.

Метод Наименьших Квадратов (МНК). Аналитическое решение

$$\sum_{i=1}^{N}(y_i - a - bx_i)^2$$ **—> MIN**

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

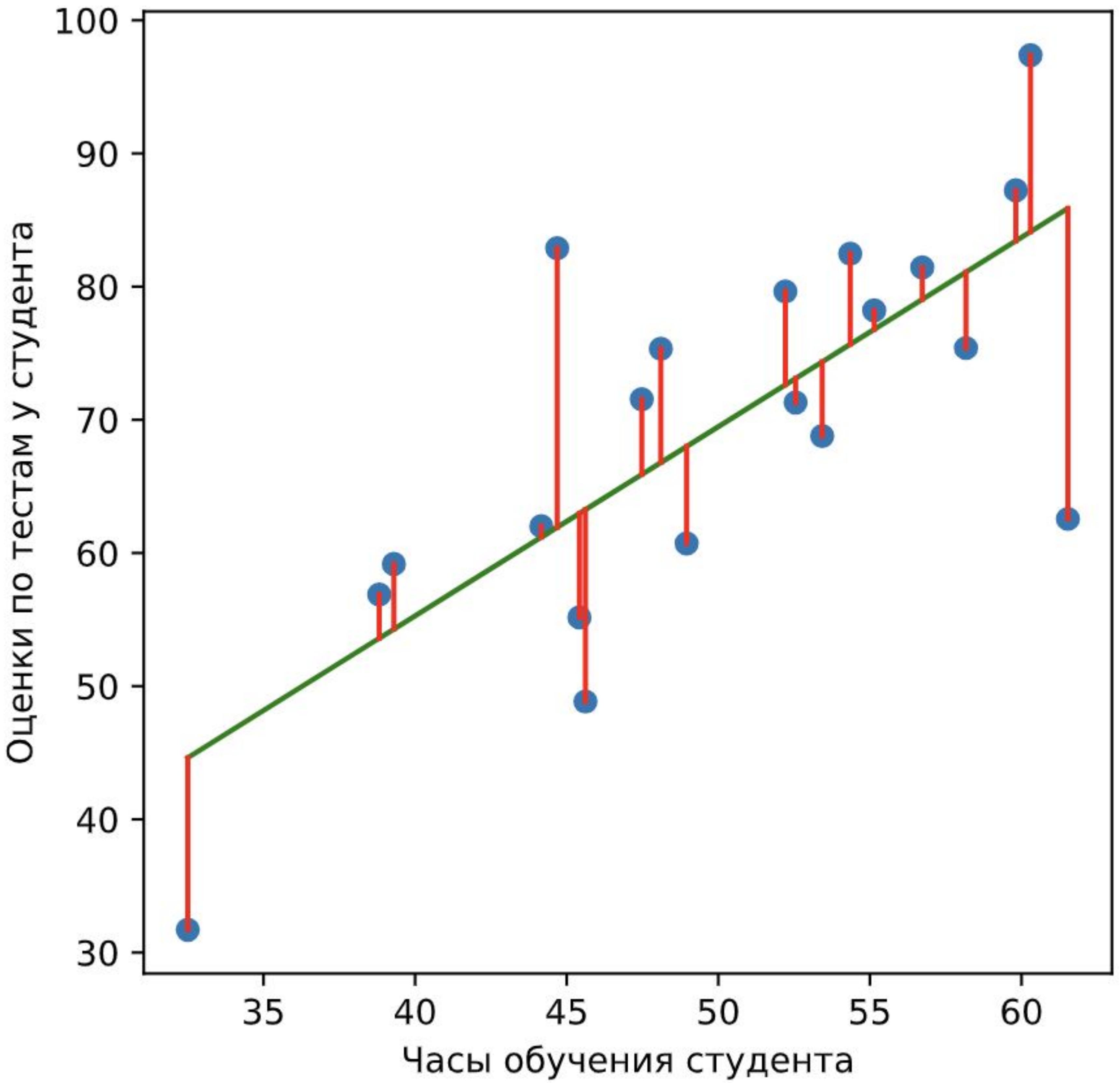
где: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Для минимизации суммы квадратов ошибок S мы берем *частные производные* по a и b и приравниваем их к нулю:

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - b\cdot x_i) = 0$$

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}x_i\cdot(y_i - a - b\cdot x_i) = 0$$

Парная линейная регрессия



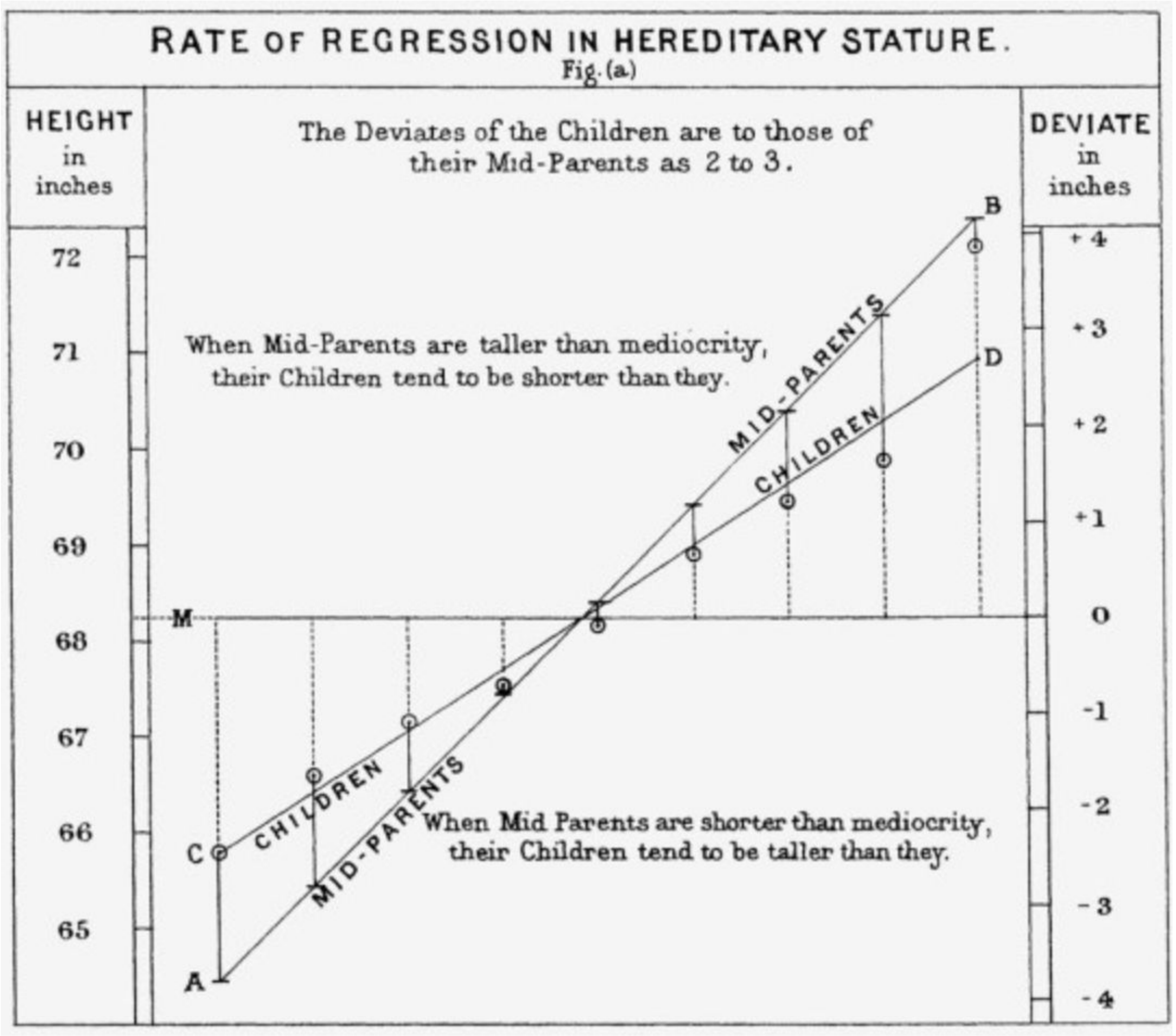
Построим линейную регрессию

сумма всех красных отрезков по модулю равна 7.88,

то есть

- MAE = 7.88,
- MSE = 98.58,
- RMSE = 9.93

Линейная регрессия. История



Почему же регрессия?

В 1886 году Понятие регрессии ввел сэр Френсис Гальтон, английский исследователь широкого профиля.

http://log-lessons.ru/chto-takoe-regressiya-k-srednemu/

Линейная регрессия. Общий случай

пусть дано *d* переменных(столбцов) x_i, составим линейную комбинацию:

$$w_1 * x_1 + w_2 * x_2 + \dots w_d * x_d$$

Линейная регрессия: $a(x) = w_0 + w_1 * x_1 + w_2 * x_2 + \dots w_d * x_d$

В компактном виде $a(x) = w_0 + \sum_{j=1}^{d} w_j x_j.$

w_i - веса/коэффициенты
w_0 - свободным коэффициентом/сдвиг

можно еще в более компактном виде $a(x) = w_0 + \langle w, x \rangle,$

а если предположить, что существует еще один столбец со всеми 1, то можно записать еще компактнее $a(x) = \langle w, x \rangle.$

Обучение линейной регрессии

Минимизируем среднеквадратическую ошибку $\frac{1}{\ell}\sum_{i=1}^{\ell}\left(\langle w, x_i\rangle - y_i\right)^2 \to \min_w$

Если продифференцировать функционал и приравнять к 0, то получим решение

$$w = (X^T X)^{-1} X^T y.$$

Но на практике так никто не делает, на практике используют численные методы,
в частности используют метод градиентного спуска (и его модификации)

Линейная регрессия. Теорема Гаусса-Маркова

Если данные обладают следующими свойствами:

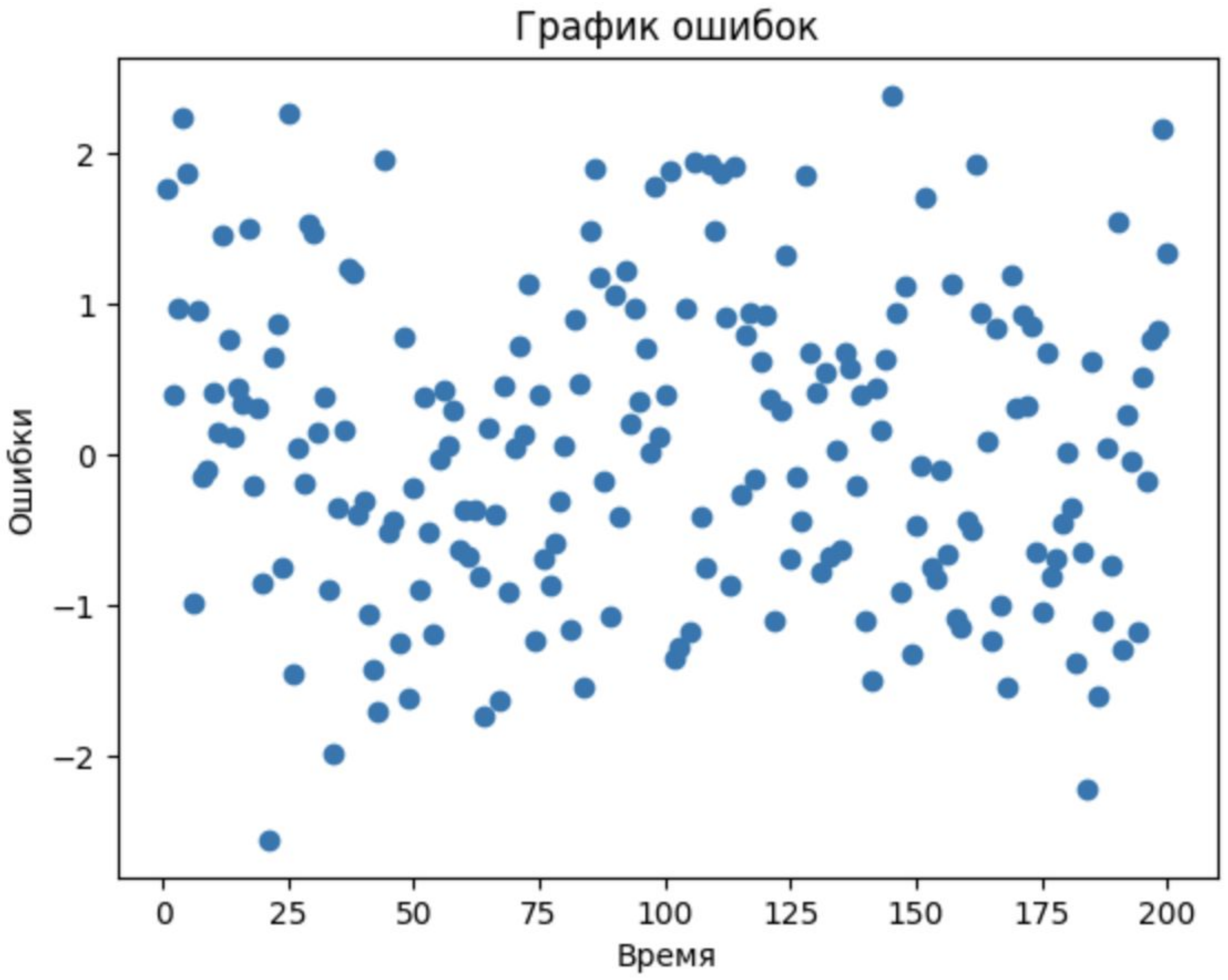
1. Модель данных правильно специфицирована;
2. Все X_i детерминированы и не все равны между собой;
3. Ошибки не носят систематического характера, то есть $\mathbb{E}(\varepsilon_i \mid X_i) = 0 \ \forall i$;
4. Дисперсия ошибок одинакова и равна некоторой σ^2;
5. Ошибки некоррелированы, то есть $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i, j$;

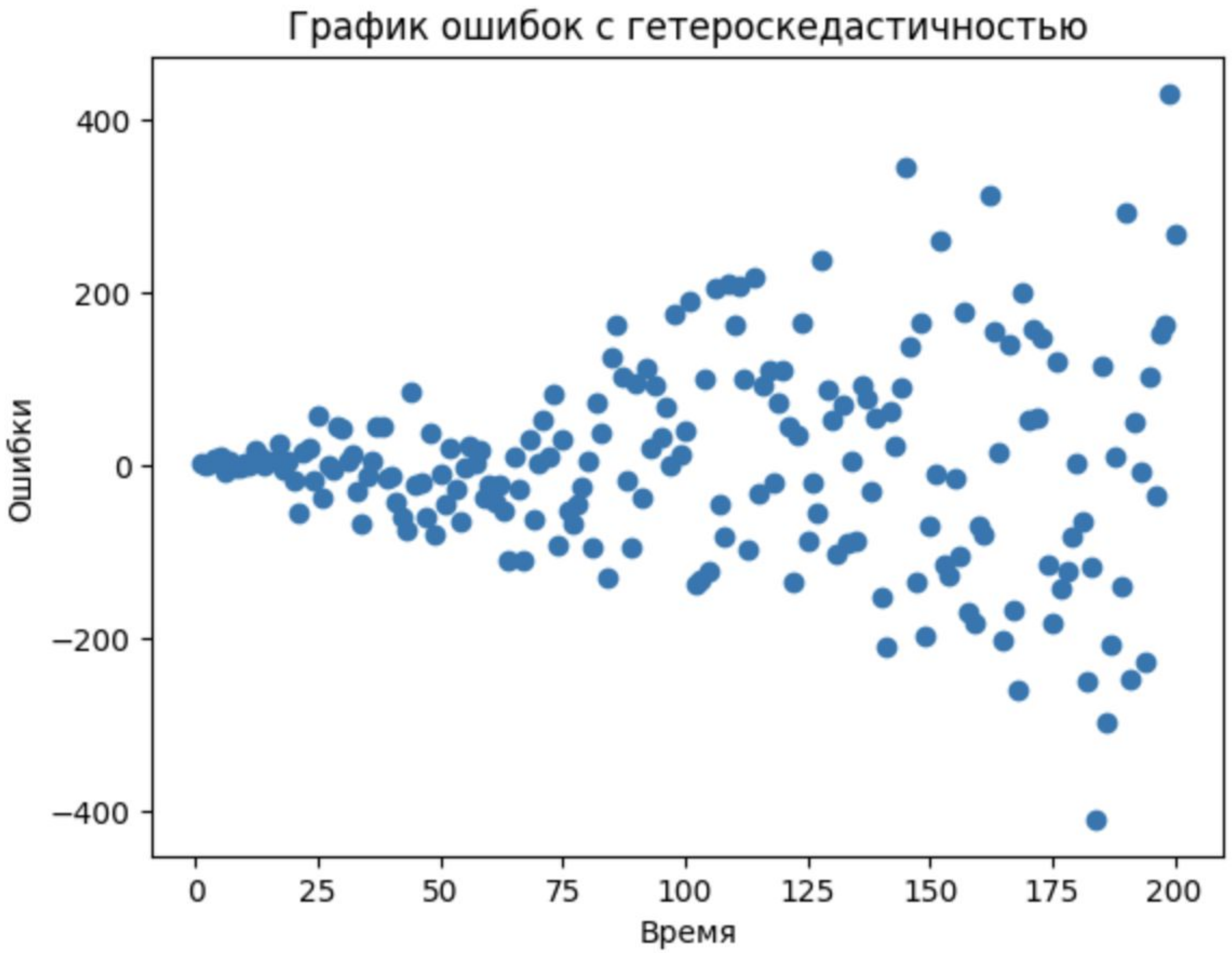
в этих условиях оценки метода наименьших квадратов оптимальны в классе линейных несмещённых оценок,

Проще говоря: метод наименьших квадратов даёт самые точные и справедливые оценки

Линейная регрессия. Теорема Гаусса-Маркова

Независимость ошибок: Ошибки должны быть независимыми и одинаково распределенными и иметь постоянную дисперсию (гомоскедастичность).

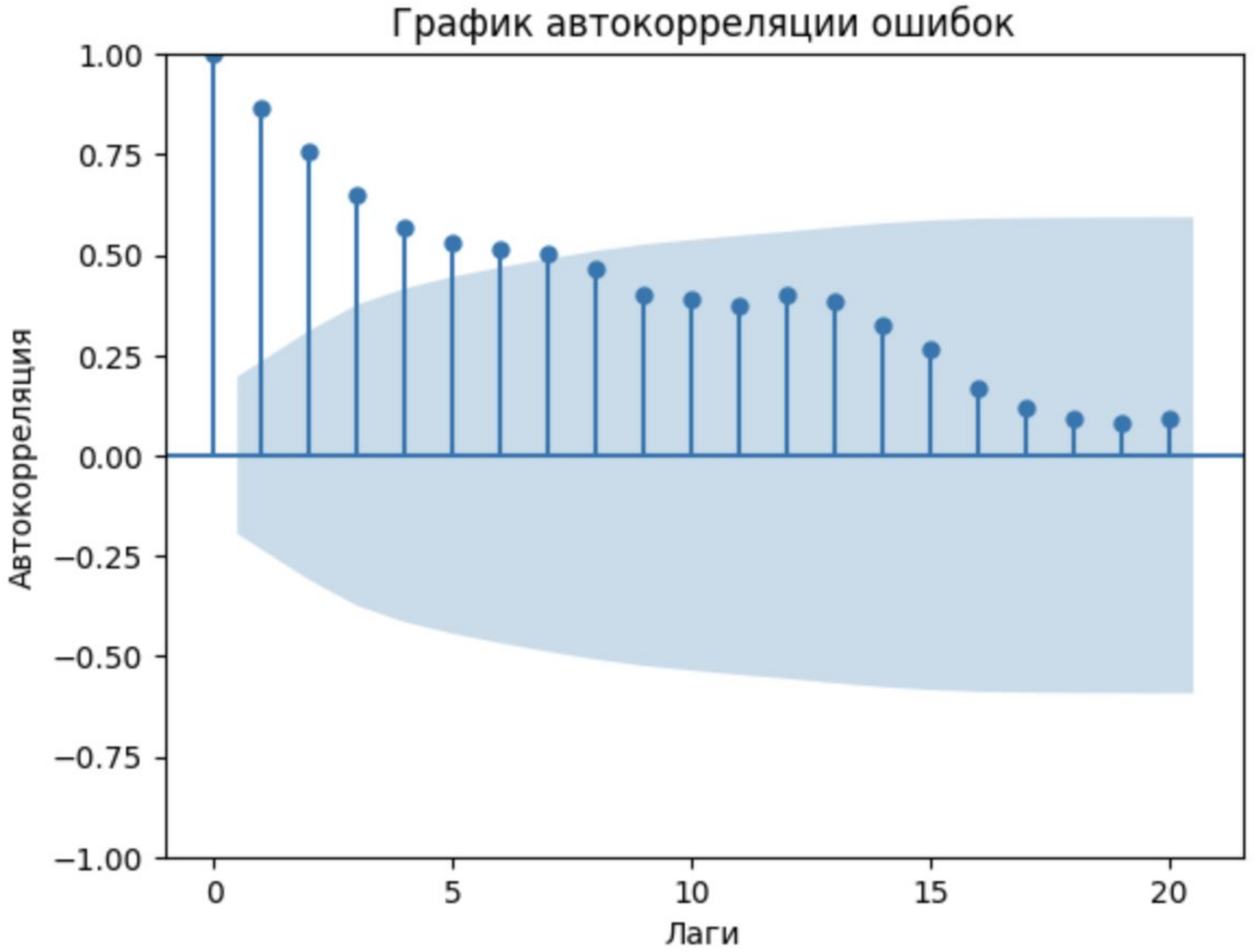


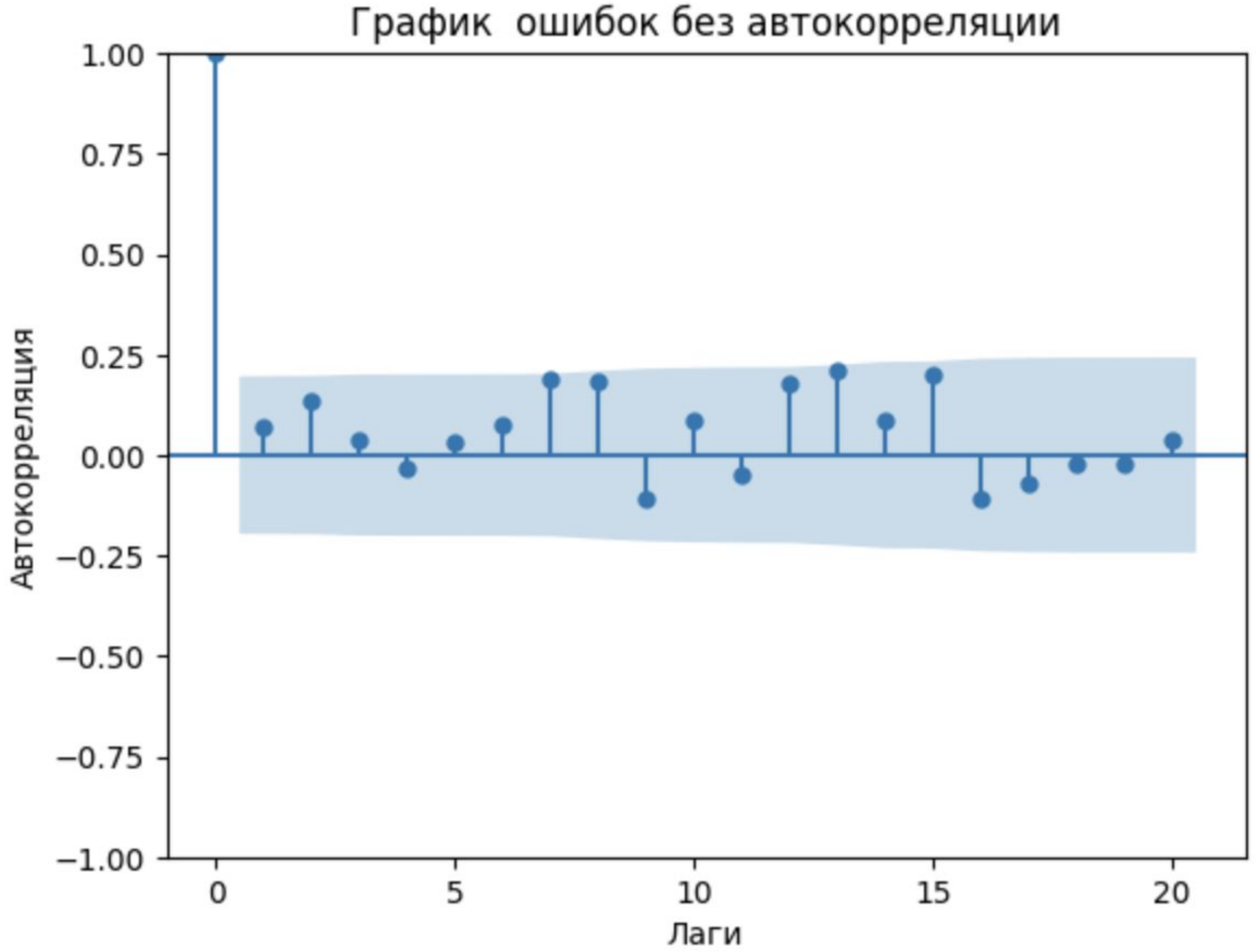


Линейная регрессия. Теорема Гаусса-Маркова

Отсутствие автокорреляции ошибок: Ошибки должны быть некоррелированными между собой.

Автокорреляция считается так: $\rho(\tau) = \mathrm{corr}(x(t), x(t+\tau)) = \frac{\mathrm{cov}(x(t), x(t+\tau))}{\sqrt{D(x(t))}\sqrt{D(x(t+\tau))}}$





УНИВЕРСИТЕТ
ИННОПОЛИС

ВОПРОСЫ И ОТВЕТЫ