Phystech@DataScience

Блок 2: линейные модели

Линейная регрессия

Модель линейной регрессии

Метод наименьших квадратов

Метрики качества в задаче регрессии

Пример

Пусть x — рост песика, а y — его вес.

Что мы знаем?

- чем крупнее песик, тем больший вес он имеет;
- песики одинакового роста могут иметь разный вес.

Выводы:

- для фиксированного роста песика x

  его вес $y = f(x)$ является *случайной величиной*;
- в среднем вес $f(x)$ *возрастает* при увеличении роста песика x.

Пример

Простая зависимость:

$$y = \theta_0 + \theta_1 x + \varepsilon,$$

x — рост песика,

y — вес песика,

θ_0, θ_1 — неизвестные параметры,

ε — случайная составляющая с *нулевым* средним.

Зависимость **линейна по параметрам**, линейна по аргументу.

Пример

Более сложная зависимость:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 + \varepsilon,$$

x_1 — рост песика,

x_2 — обхват туловища песика,

y — вес песика,

$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры,

ε — случайная составляющая с *нулевым* средним.

Зависимость **линейна по параметрам**,

но квадратична по аргументам.

Модель линейной регрессии

Рассматриваем функциональную зависимость вида

$$y = y(x) = \theta_1 x_1 + ... + \theta_d x_d$$

$x_1, ..., x_d$ — признаки,

$\theta = (\theta_1, ..., \theta_d)^T$ — вектор параметров.

Для оценки θ производится n испытаний вида

$$Y_i = \theta_1 x_{i1} + ... + \theta_d x_{id} + \varepsilon_i, \quad i = 1, ..., n,$$

$x_i = (x_{i1}, ..., x_{id})$ — признаковые описания объекта i (обычно неслучайные),

$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ — случайная ошибка измерений.

Модель линейной регрессии

Введем обозначения

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Матричная форма записи проведенных испытаний:

$$Y = X\theta + \varepsilon.$$

$X \in \mathbb{R}^{n \times d}$ — регрессоры (или матрица плана эксперимента),

$Y \in \mathbb{R}^n$ — отклик.

Матричный вид зависимости: $y(x) = x^T\theta$.

Замечание

Зависимость $y = y(x)$ должна быть **линейна по параметрам**,

но ***не*** *обязана* быть линейной *по признакам*.

Пусть $z_1, ..., z_k$ — набор "независимых" переменных.

Можно рассматривать модель

$$y(x) = \theta_1 x_1(z_1, ..., z_k) + ... + \theta_d x_d(z_1, ..., z_k),$$

где $x_j(z_1, ..., z_k)$ — некоторые функции (м.б. нелинейные).

Примеры:

- $x(z_1, ..., z_k) = 1$;
- $x(z_1, ..., z_k) = z_1$;
- $x(z_1, ..., z_k) = \ln z_1$;
- $x(z_1, ..., z_k) = z_1^2 z_2$.

Пример: Потребление мороженого

Предполагается *линейная зависимость* потребления мороженого в литрах на человека от среднесуточной температуры воздуха: $ic = \theta_0 + \theta_1 t$.

Проведена серия наблюдений

$$IC_i = \theta_0 + \theta_1 t_i + \varepsilon_i,$$

t_i — среднесуточная температура воздуха,

IC_i — потребление мороженого в литрах на чел.,

$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ — случайное отклонение.

Пример: Потребление мороженого

Наблюдения: $IC_i = \theta_0 + \theta_1 t_i + \varepsilon_i$.

В данном примере $x_1(t) = 1, x_2(t) = t$,

$$X = \begin{pmatrix} 1 & t_1 \\ \dots & \\ 1 & t_n \end{pmatrix}, \; Y = \begin{pmatrix} IC_1 \\ \dots \\ IC_n \end{pmatrix}, \; \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}.$$

Пусть $w = I\{\text{выходной день}\}$, зависимость $ic = \theta_0 + \theta_1 t + \theta_2 t^2 w$.

Наблюдения: $IC_i = \theta_0 + \theta_1 t_i + \theta_2 t_i^2 w_i + \varepsilon_i$.

В данном примере $x_1(t, w) = 1, x_2(t, w) = t, x_3(t, w) = t^2 w$,

$$X = \begin{pmatrix} 1 & t_1 & t_1^2 w_1 \\ \dots & & \\ 1 & t_n & t_n^2 w_n \end{pmatrix}, \; Y = \begin{pmatrix} IC_1 \\ \dots \\ IC_n \end{pmatrix}, \; \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}.$$

Модель линейной регрессии

Метод наименьших квадратов

Метрики качества в задаче регрессии

Метод наименьших квадратов

Зависимость: $y(x) = x^T\theta$, $\theta \in \mathbb{R}^d$.

Испытания: $Y = X\theta + \varepsilon$, $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^n$.

Хотим как-то оценить параметр θ на основании полученных данных.

Пусть $\widehat{\theta} = \widehat{\theta}(X, Y)$ — наша оценка θ.

Как понять, что она хорошая?

Метрика mean squared error (MSE):

$$MSE(\widehat{\theta}) = \left|\left|Y - X\widehat{\theta}\right|\right|^2$$

Оценка $\widehat{\theta} = \arg\min_{\theta} MSE(\widehat{\theta})$ называется оценкой по методу наименьших квадратов параметра θ.

Метод наименьших квадратов

Теорема. Если матрица X^TX невырождена, то $\widehat{\theta} = (X^TX)^{-1}X^TY$.

$$MSE(\theta) = ||Y - X\theta||^2 = (Y - X\theta)^T(Y - X\theta) = Y^TY - 2Y^TX\theta + \theta^TX^TX\theta$$

Берем производную по θ и приравниваем ее к нулю:

$$\frac{\partial MSE(\theta)}{\partial \theta} = -2Y^TX + 2\theta^TX^TX = 0$$

Отсюда получается утверждение теоремы. □

Предсказанием *отклика* на новом объекте x будет величина $\widehat{y(x)} = x^T\widehat{\theta}$.

Некоторые свойства (для справки)

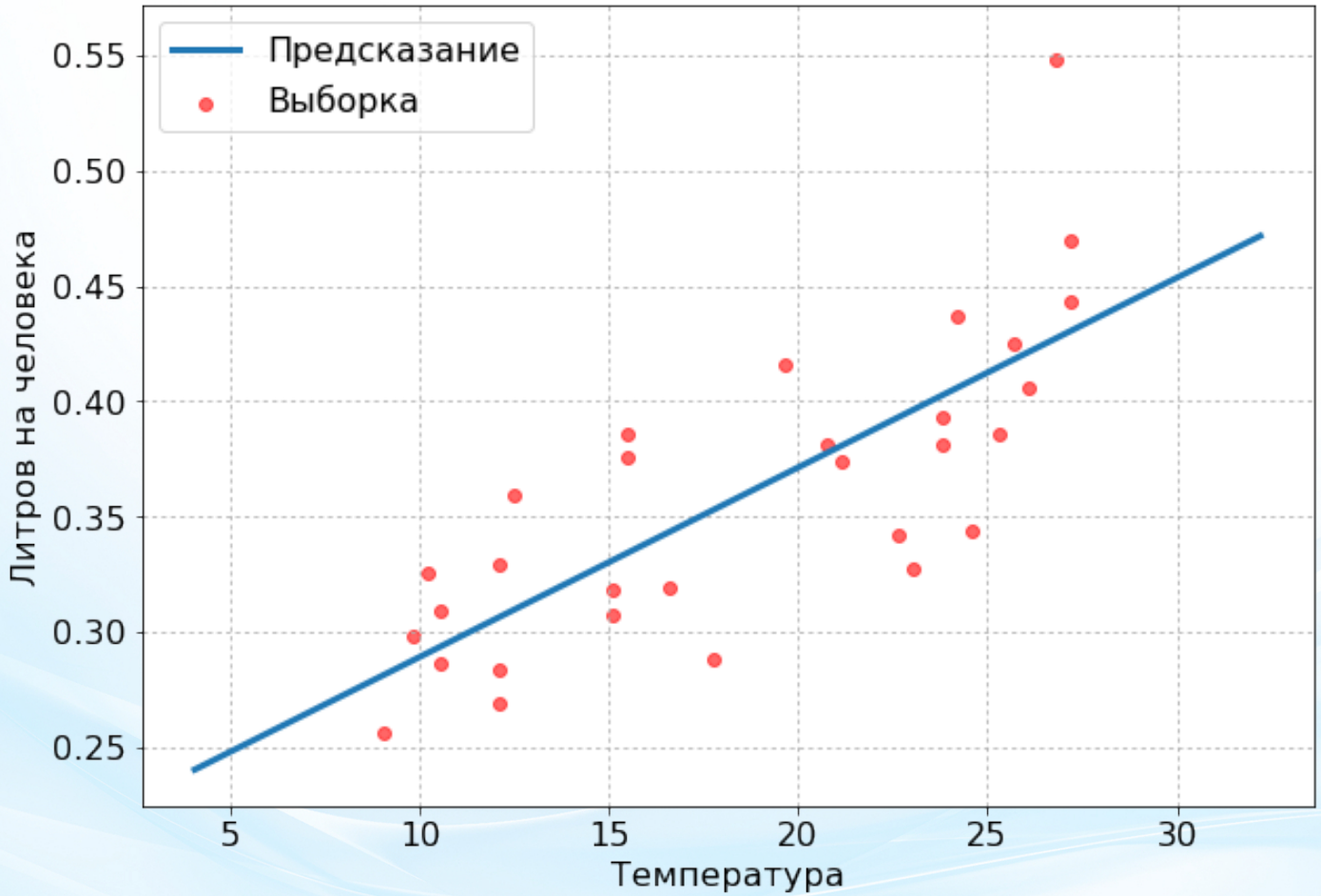
Если выполнено:

1. $rkX = d$ (признаки линейно-независимы)
2. для $\varepsilon = y - \widehat{y}$
   - ε_i одинаково распределены
   - $\mathrm{E}\varepsilon_i = 0$
   - $\mathrm{D}\varepsilon_i = \sigma^2$
   - $\mathrm{E}\varepsilon_i\varepsilon_j = 0, \forall i \neq j$

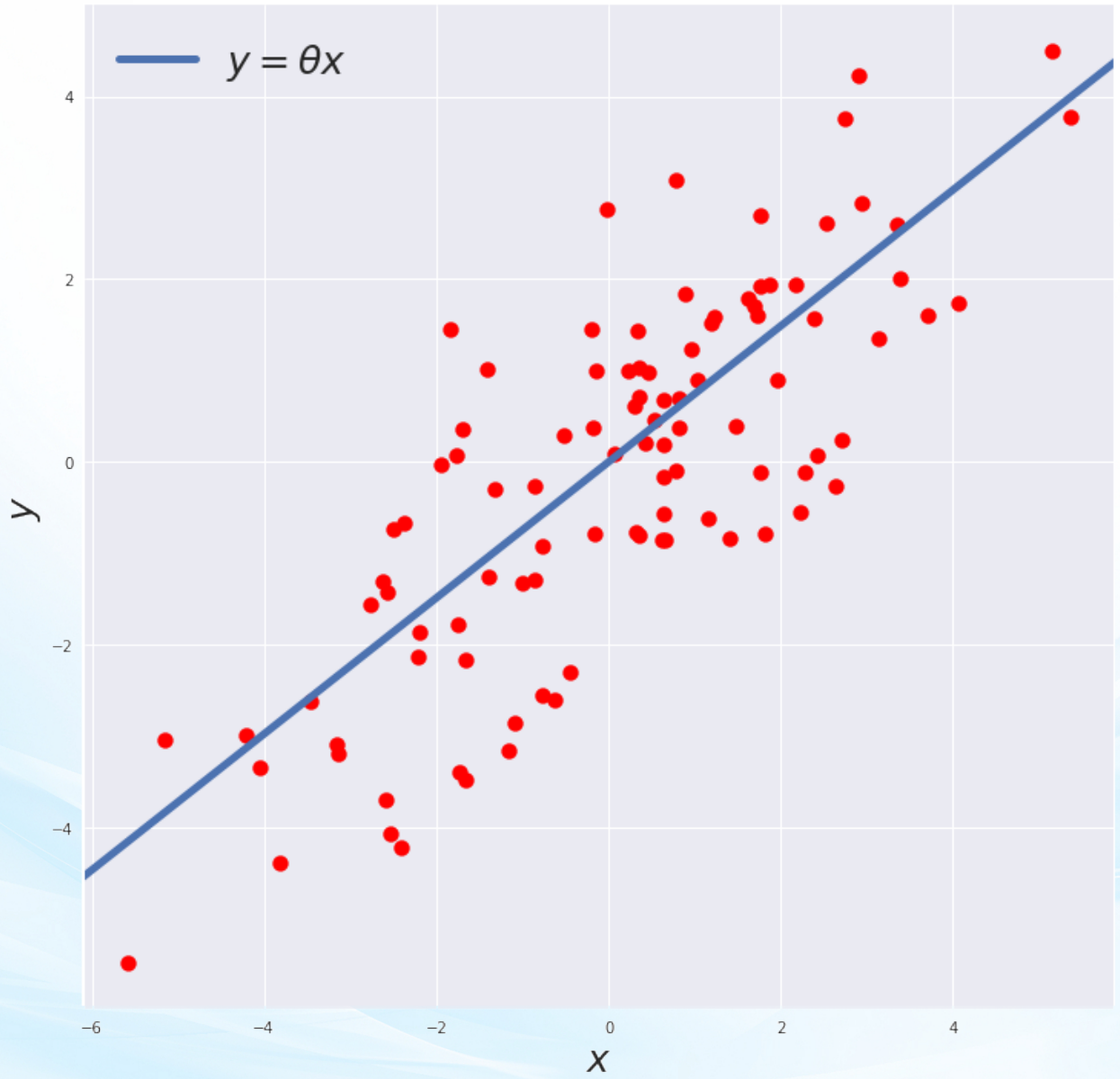
Тогда оценка МНК — хорошая:

1. Несмещённая: $\mathrm{E}\widehat{\theta} = \theta$
2. $\mathrm{D}\widehat{\theta} = \sigma^2(X^TX)^{-1}$
3. Эффективная

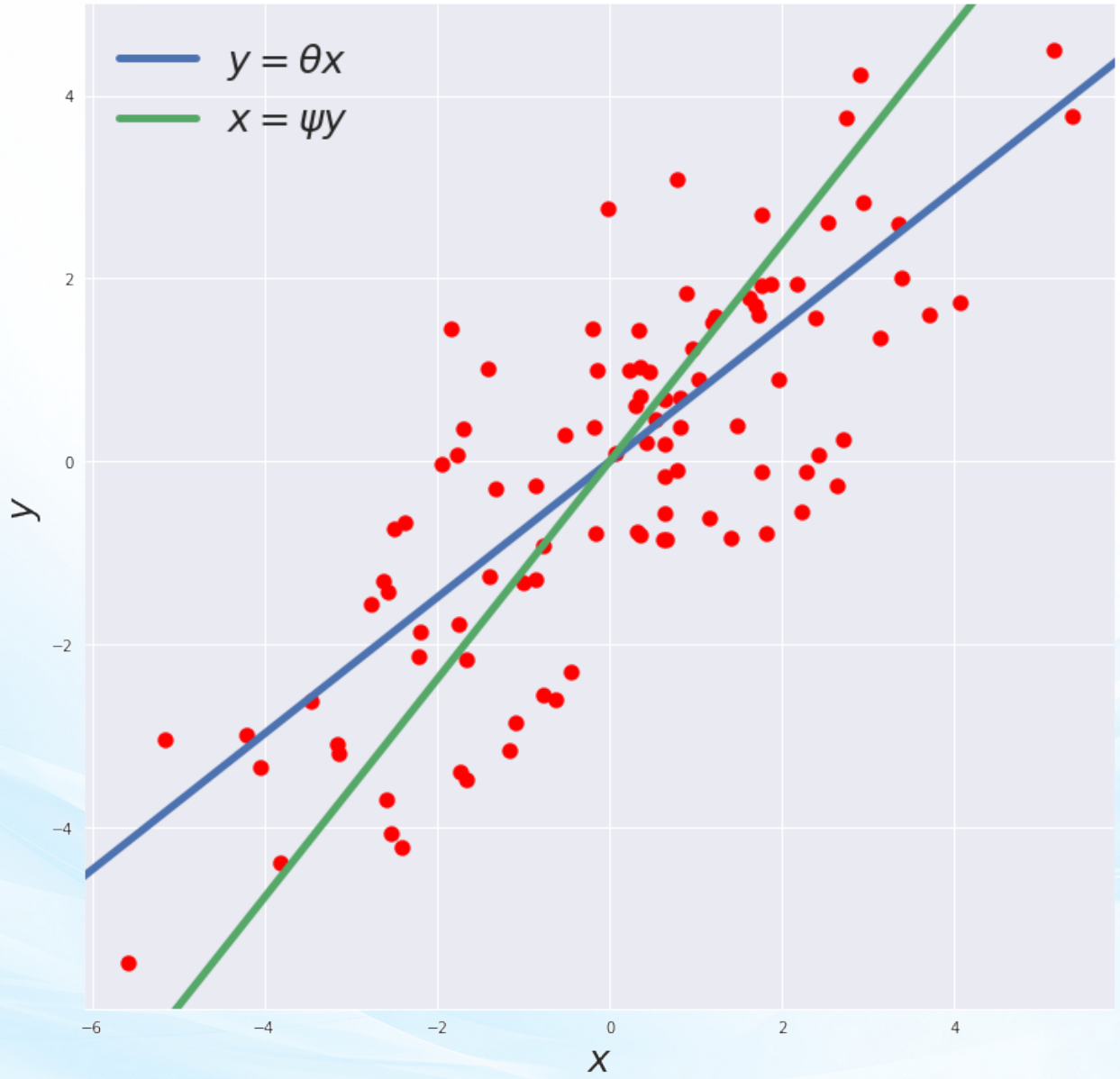
Пример: Потребление мороженого



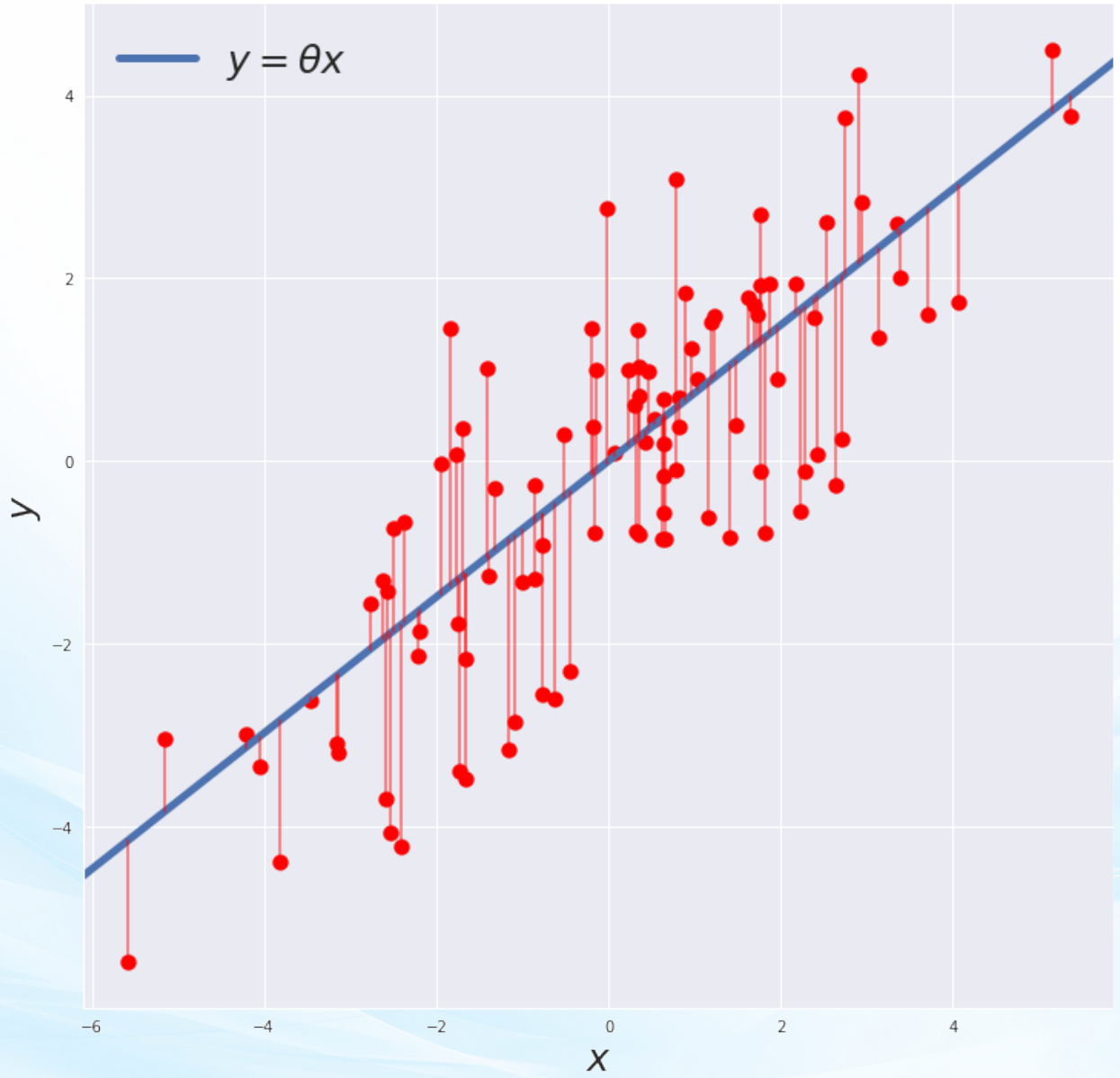
Инверсия



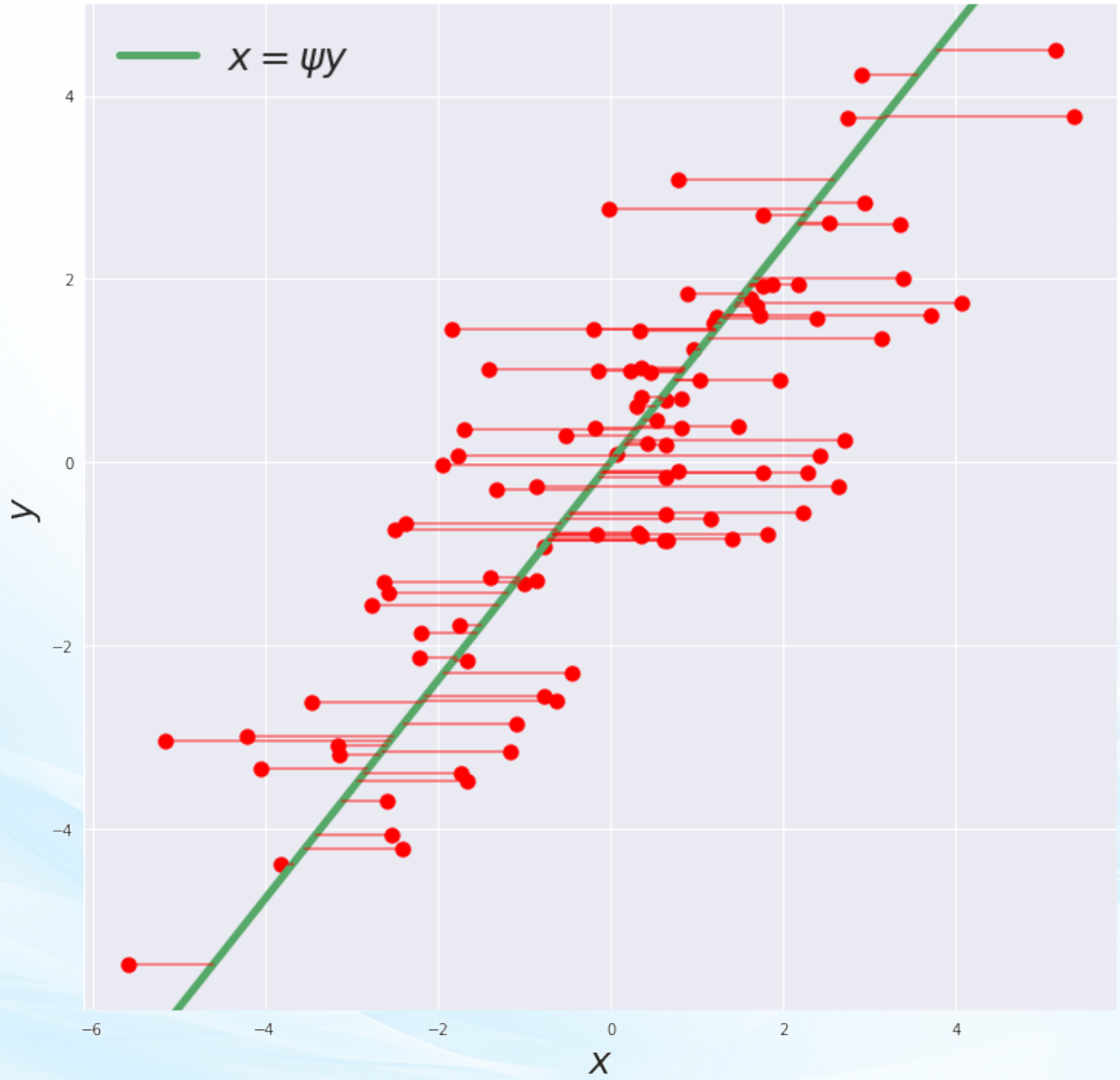
Инверсия



Инверсия



Инверсия



Категориальные переменные

x — id астрономического объекта (натуральное число),

y — его масса.

Предположим, что типы занумерованы следующим образом:

- $x = 1$ — *черная дыра;*
- $x = 2$ — *нейтронная звезда;*
- $x = 3$ — *обычная звезда.*

	id	тип	масса
0	2546	2	1.1
1	3642	1	30.0
2	1211	3	5.5
3	4333	3	0.7

Если $x \in \{1, ..., k\}$, то рассматриваются **dummy-переменные**:

$$x_j = I\{x = j\}, \quad j = 1, ..., k-1,$$

модель $y = \theta_0 + \theta_1 x_1 + ... + \theta_{k-1} x_{k-1}$.

	id	ч. дыра	нейт. зв.	масса
0	2546	0	1	1.1
1	3642	1	0	30.0
2	1211	0	0	5.5
3	4333	0	0	0.7

Модель линейной регрессии

Метод наименьших квадратов

Метрики качества в задаче регрессии

Метрики качества в задаче регрессии

Y — реальные наблюдения, $\widehat{Y}$ — предсказания.

- Mean Squared Error:

$$MSE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \quad - \quad \text{легко оптимизируется}$$

- Mean Absolute Error:

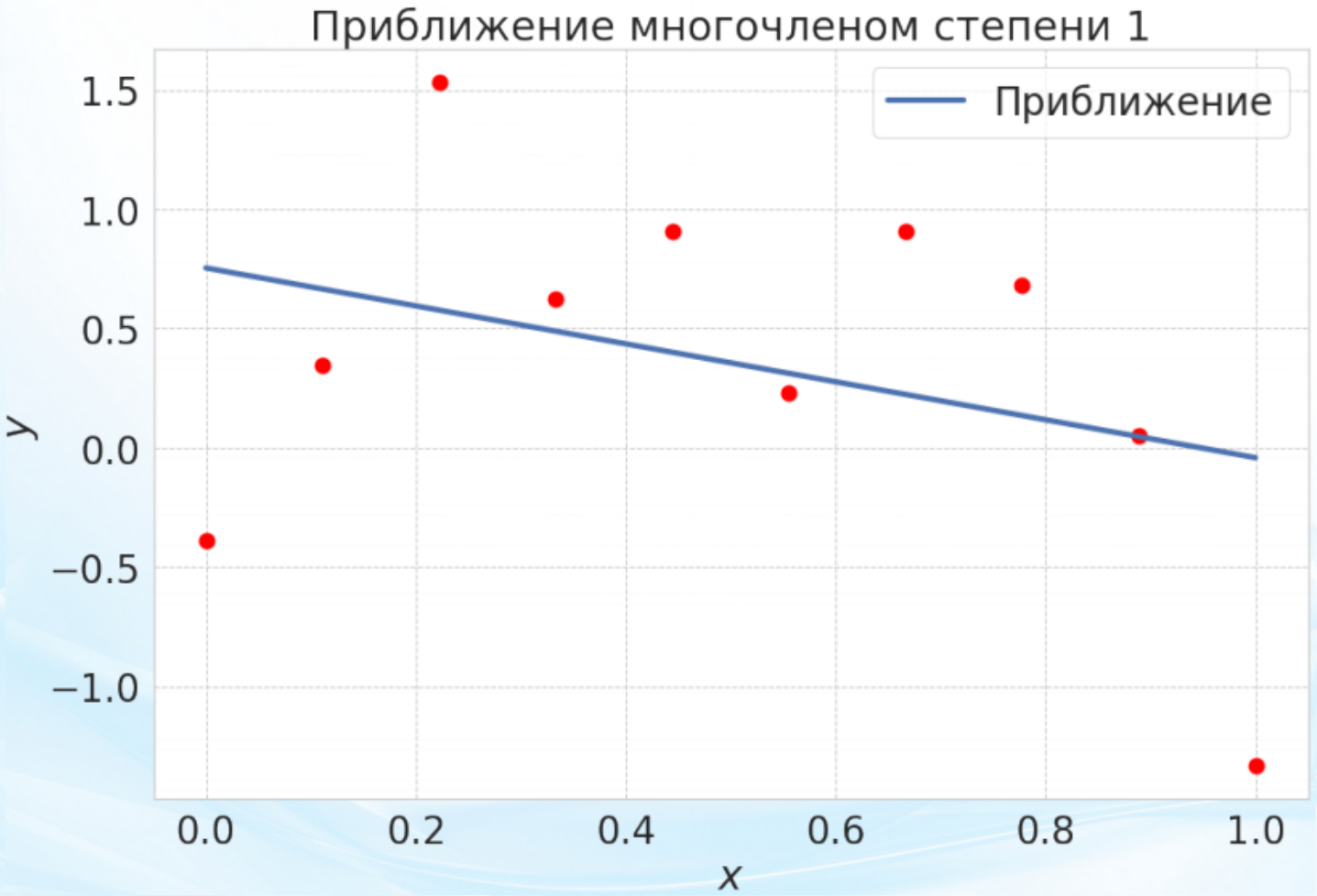
$$MAE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i| \quad - \quad \text{устойчивее к выбросам}$$

- Mean Absolute Percentage Error:

$$MAPE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| * 100\% \quad - \quad \text{информативна}$$

Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность



Недообучение

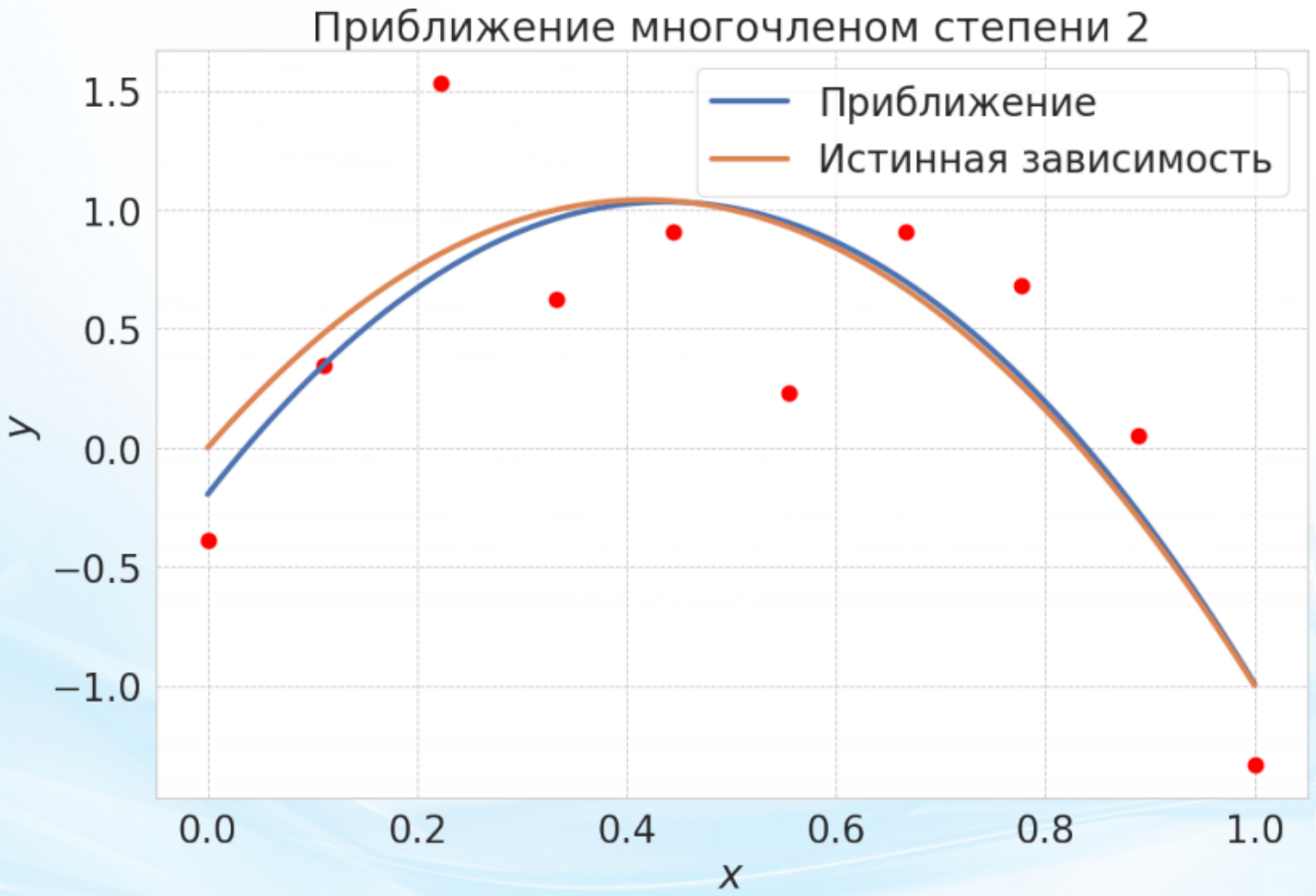
Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность

Переобучение

Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность



Нормально!

Тренировочная и тестовая выборки

Если все время работать *с одной и той же* выборкой
(это жаргон, корректно понимать "реализацией выборки")
и все больше улучшать модель, "подгонять" ее под выборку,

может возникнуть *переобучение*.

Предсказание на новом объекте может быть неадекватным.

Поэтому перед началом работы имеющиеся данные делят на две части:

тренировочную (обучающую) и тестовую выборки.

На тренировочной выборке происходит обучение моделей (например, оценка коэффициентов в линейной регрессии).

На тестовой выборке происходит оценка качества итоговой модели с использванием метрик качества.


ВСЁ!