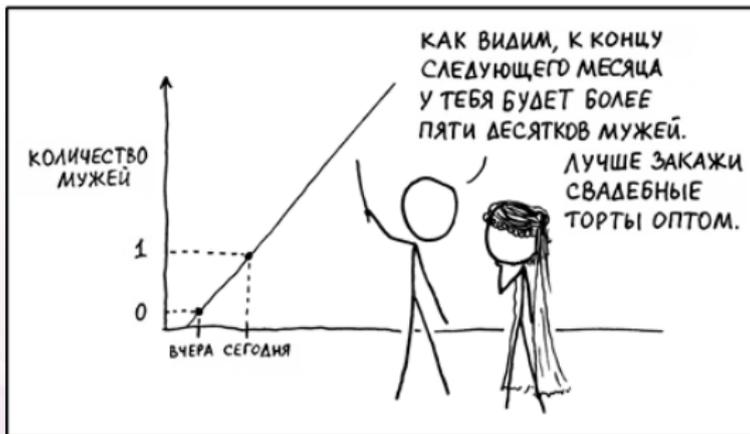




# Линейная регрессия

МОЁ ХОББИ: ЭКСТРАПОЛИРОВАТЬ





## Пример

Пусть  $x$  — рост котика, а  $y$  — его вес.

Что мы знаем?

- ▶ чем крупнее котик, тем больший вес он имеет;
- ▶ котики одинакового роста могут иметь разный вес.

Выводы:

- ▶ для фиксированного роста котика  $x$  его вес  $y = f(x)$  является случайной величиной;
- ▶ в среднем вес  $f(x)$  возрастает при увеличении роста котика  $x$ .



## Пример

Простая зависимость:

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

$x$  — рост котика,

$y$  — вес котика,

$\theta_0, \theta_1$  — неизвестные параметры,

$\varepsilon$  — случайная составляющая  
с нулевым средним.

Зависимость

- ▶ **линейна по параметрам,**
- ▶ линейна по аргументу.

Более сложная зависимость:

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

$x_1$  — рост котика,

$x_2$  — обхват туловища котика,

$y$  — вес котика,

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

$\varepsilon$  — случайная составляющая  
с нулевым средним.

Зависимость

- ▶ **линейна по параметрам,**
- ▶ квадратична по аргументам.



# Модель линейной регрессии

Рассматриваем функциональную зависимость вида

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

$x_1, \dots, x_d$  — признаки,

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

Предполагаем, что данные соотносятся с этой зависимостью с некоторым шумом

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

$x_i = (x_{i1}, \dots, x_{id})$  — признаки  $i$ -го объекта (обычно неслучайные),

$\varepsilon_i$  — случайная ошибка (погрешности).



## 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 обучающих данных

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

$X \in \mathbb{R}^{n \times d}$  — матрица признаков,

$Y \in \mathbb{R}^n$  — таргет (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  $z_1, \dots, z_k$  — набор "независимых" переменных.

Можно рассматривать модель

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

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

Примеры:

▶  $x(z_1, \dots, z_k) = 1;$

▶  $x(z_1, \dots, z_k) = z_1;$

▶  $x(z_1, \dots, z_k) = \ln z_1;$

▶  $x(z_1, \dots, z_k) = z_1^2 z_2.$



## Пример: Потребление мороженого

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

Получены данные

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

$t_i$  — среднесуточная температура воздуха,

$IC_i$  — потребление мороженого в литрах на чел.,

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





## Пример: Потребление мороженого

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

В этом примере  $x_0(t) = 1$ ,  $x_1(t) = t$ ,

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

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

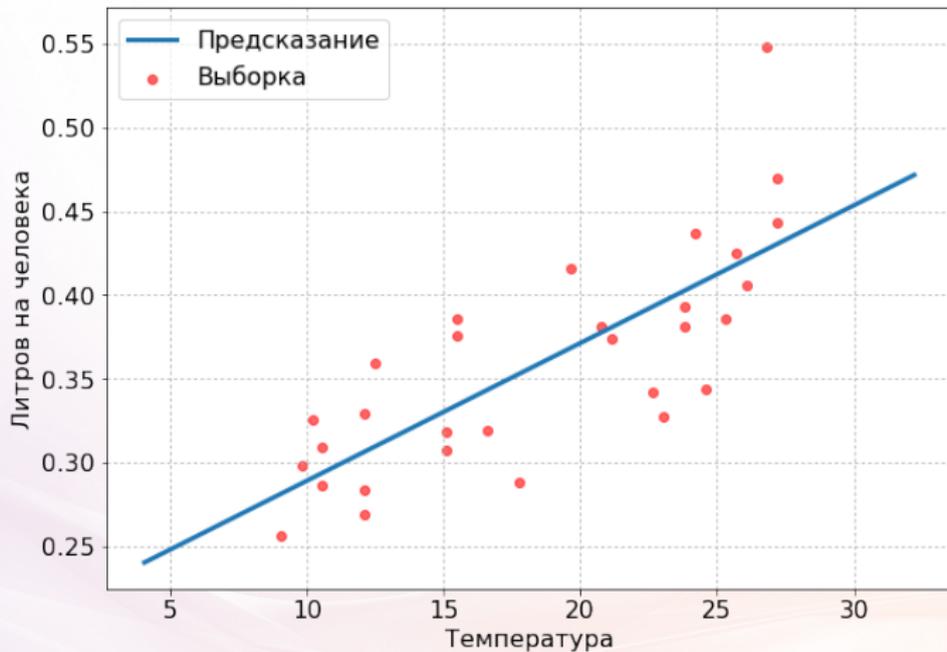
Данные:  $IC_i = \theta_0 + \theta_1 t_i + \theta_2 t_i^2 w_i + \varepsilon_i$ .

В этом примере  $x_0(t, w) = 1$ ,  $x_1(t, w) = t$ ,  $x_2(t, w) = t^2 w$ ,

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

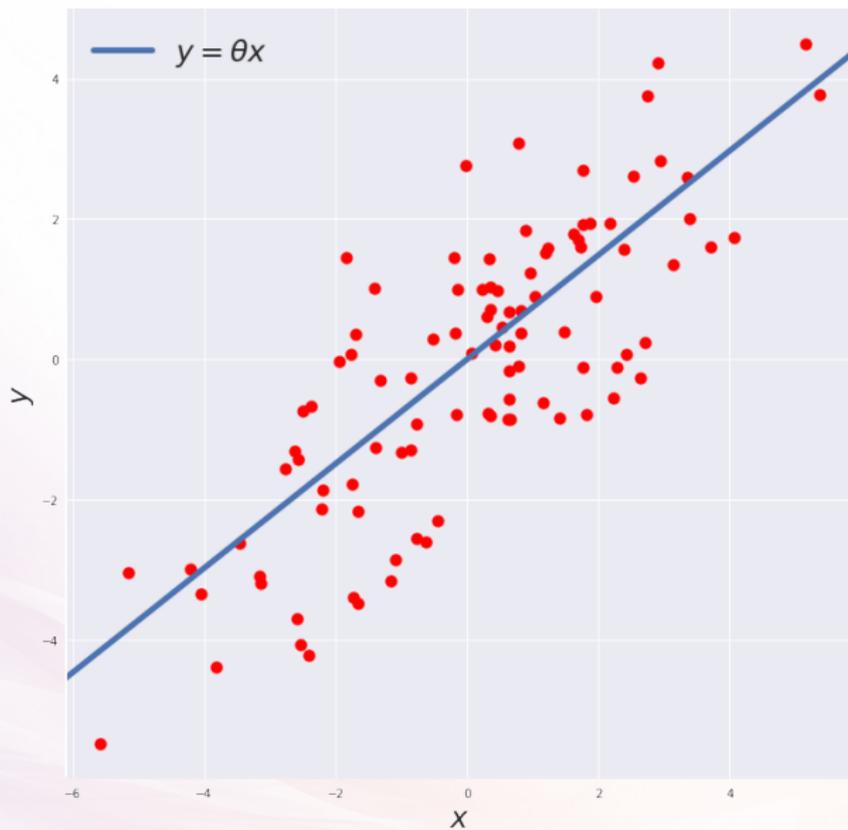


## Пример: Потребление мороженого



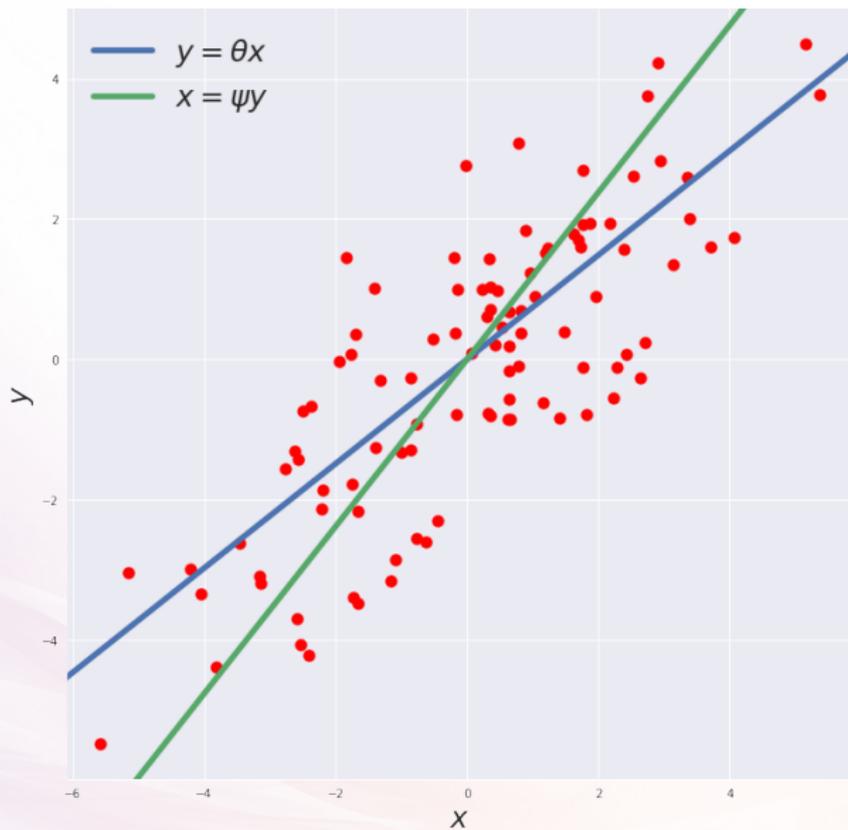


# Инверсия



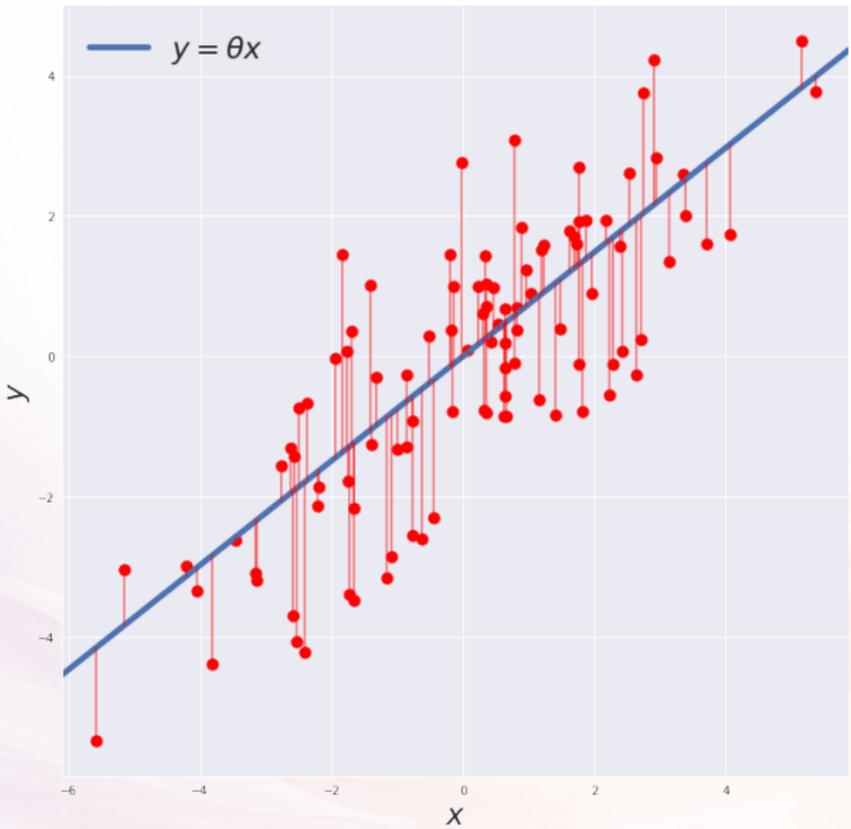


# Инверсия





# Инверсия





# Инверсия

