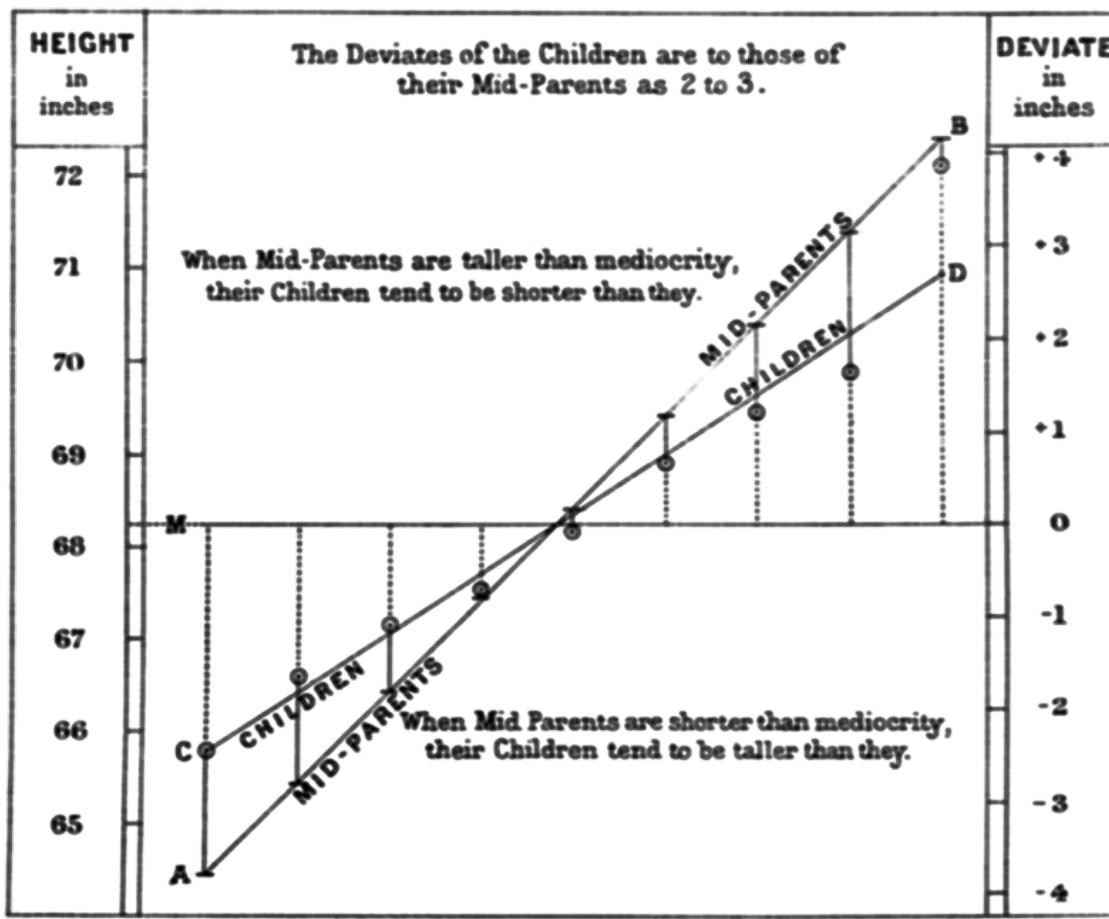


# РЕГРЕССИЯ

---

# ТЕРМИН

› Френсис Гальтон, 1885:  $y - \bar{y} \approx \frac{2}{3}(x - \bar{x})$



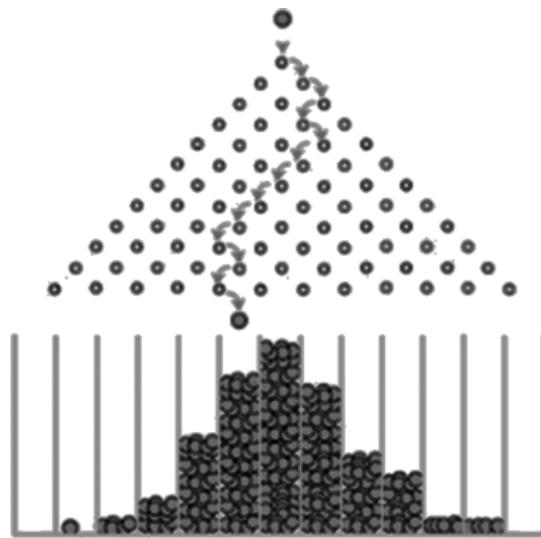
# РЕГРЕССИЯ К СРЕДНЕМУ

---

› Машина Гальтона

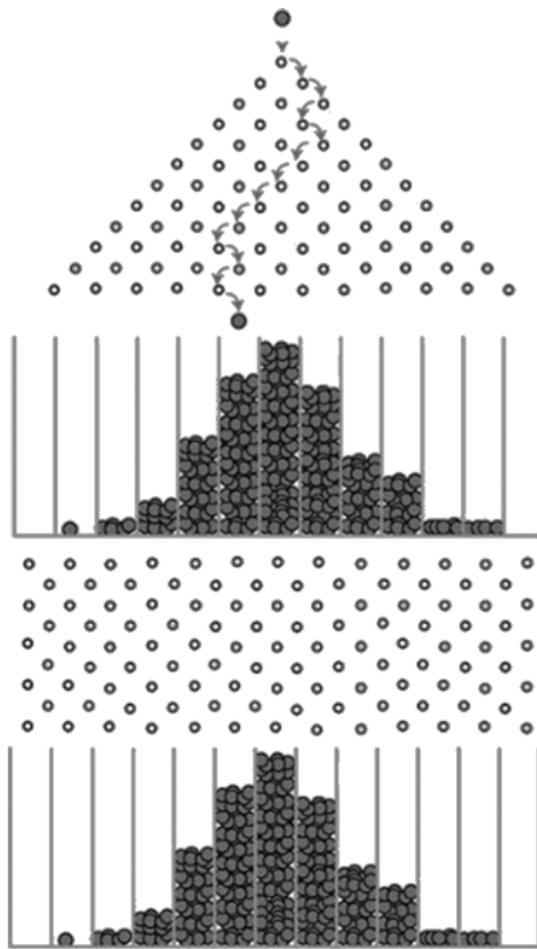
# РЕГРЕССИЯ К СРЕДНЕМУ

---



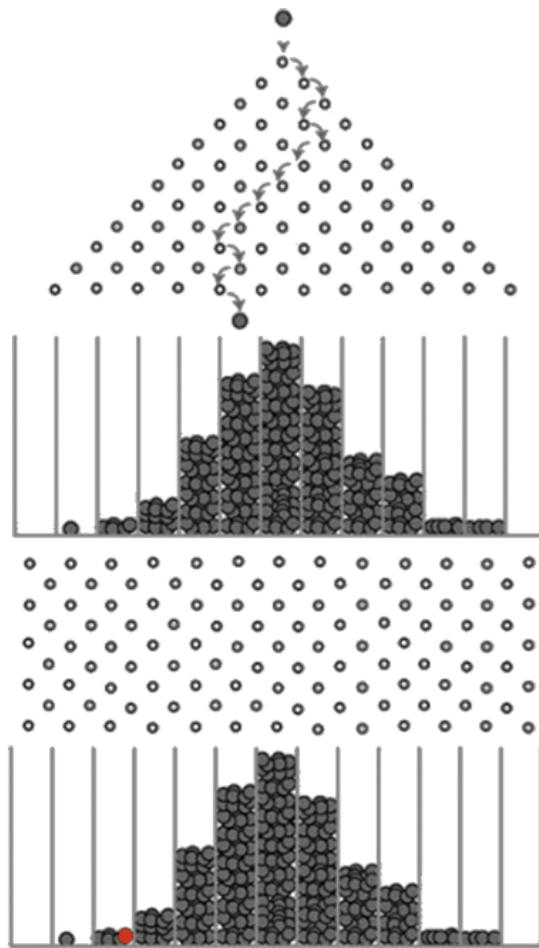
# РЕГРЕССИЯ К СРЕДНЕМУ

---



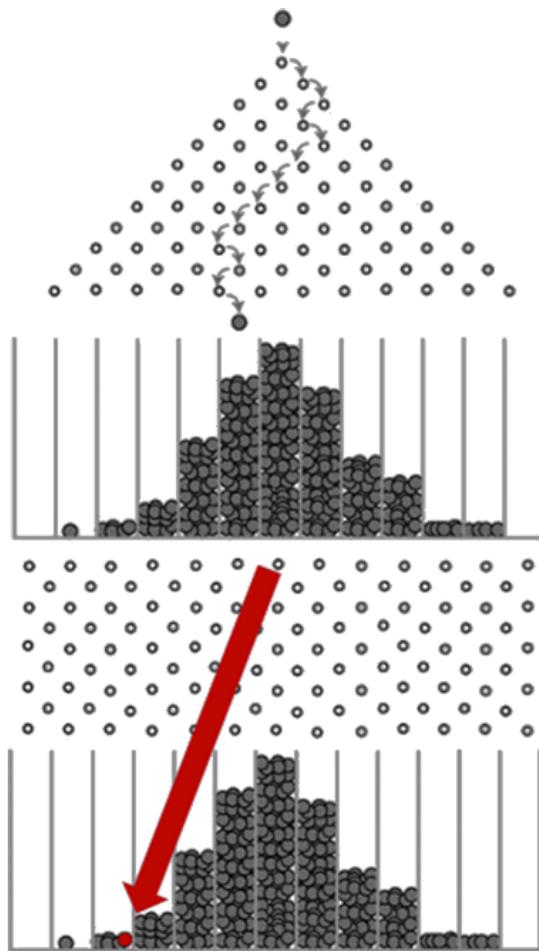
# РЕГРЕССИЯ К СРЕДНЕМУ

---



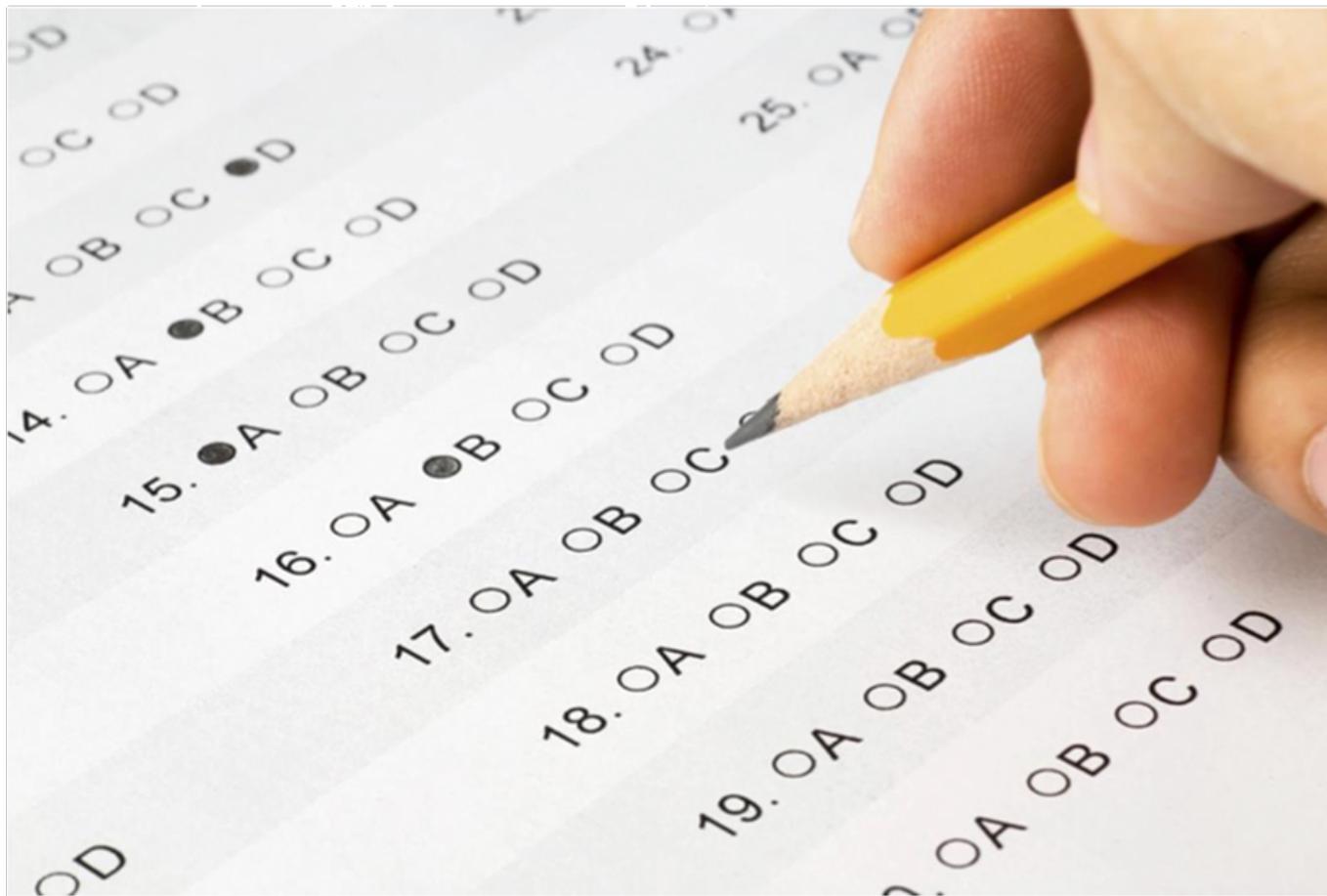
# РЕГРЕССИЯ К СРЕДНЕМУ

---



# РЕГРЕССИЯ К СРЕДНЕМУ

---



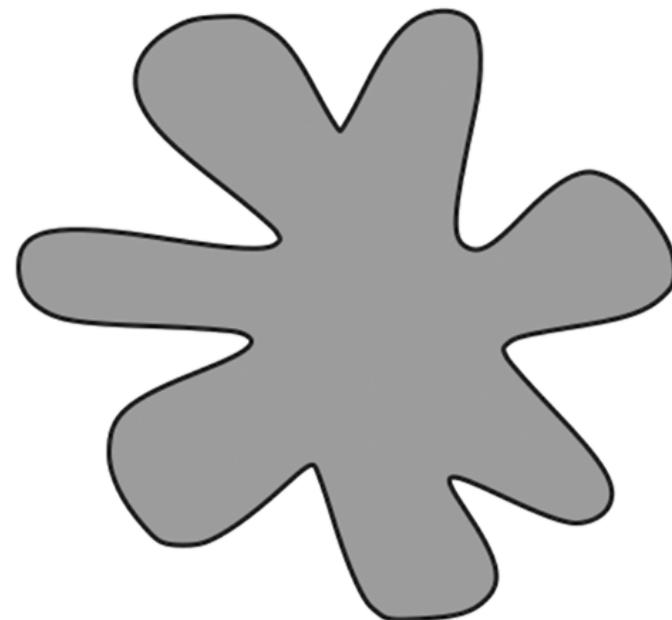
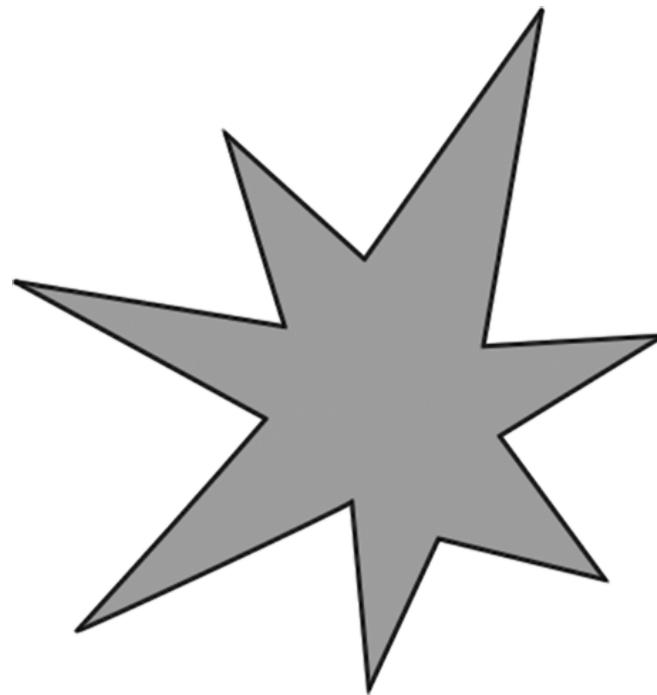
# ДРУГИЕ РАБОТЫ ГАЛЬТОНА

---



# ДРУГИЕ РАБОТЫ ГАЛЬТОНА

---



# ДРУГИЕ РАБОТЫ ГАЛЬТОНА

---



# ДРУГИЕ РАБОТЫ ГАЛЬТОНА

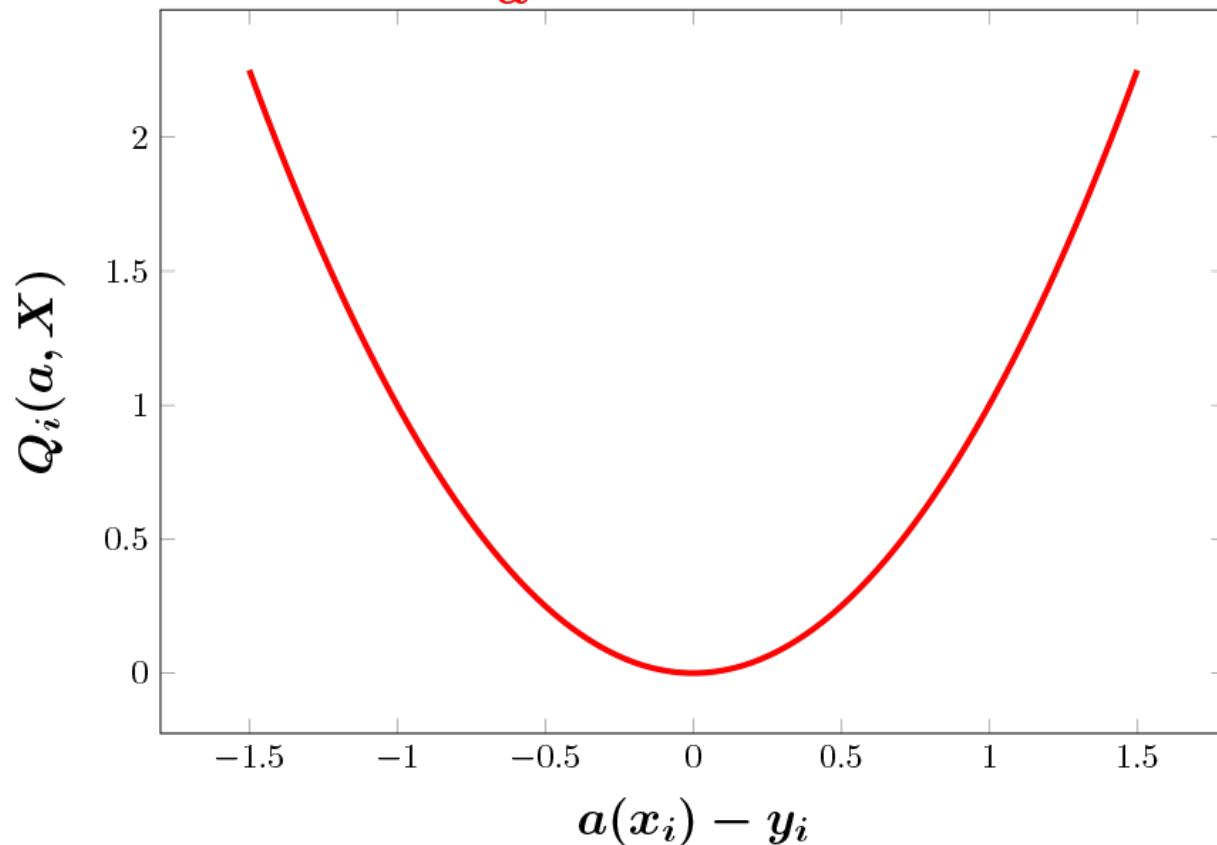
---



# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$a_*(x) = \underset{a}{\operatorname{argmin}} Q(a, X)$$

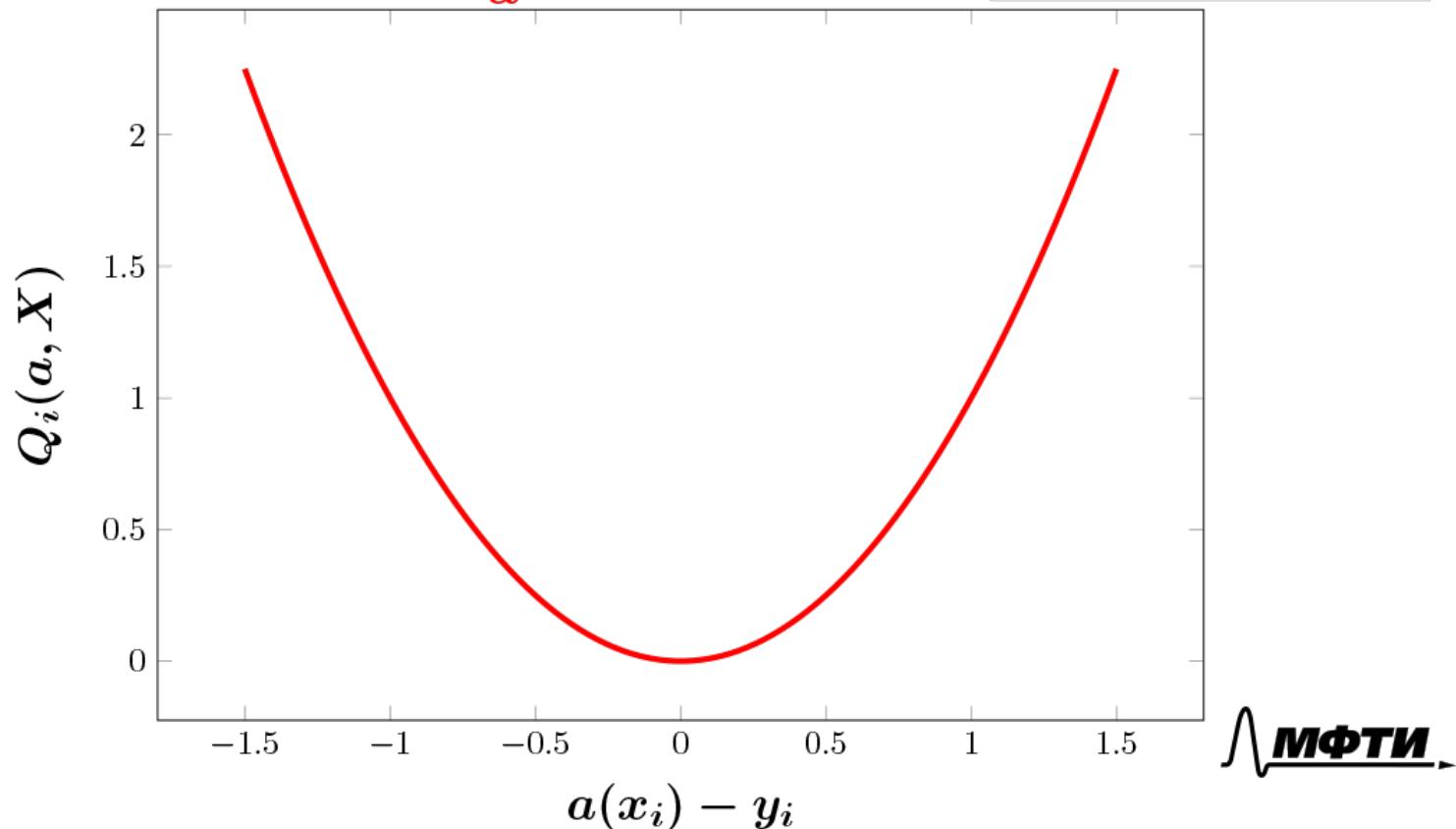


# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$a_*(x) = \underset{a}{\operatorname{argmin}} Q(a, X)$$

метод  
наименьших  
квадратов



# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

» Линейная регрессия:

$$Q(\mathbf{w}, \mathbf{X}) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

$$\mathbf{w}_*(x) = \underset{\mathbf{w}}{\operatorname{argmin}} Q(\mathbf{w}, \mathbf{X})$$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

» Линейная регрессия:

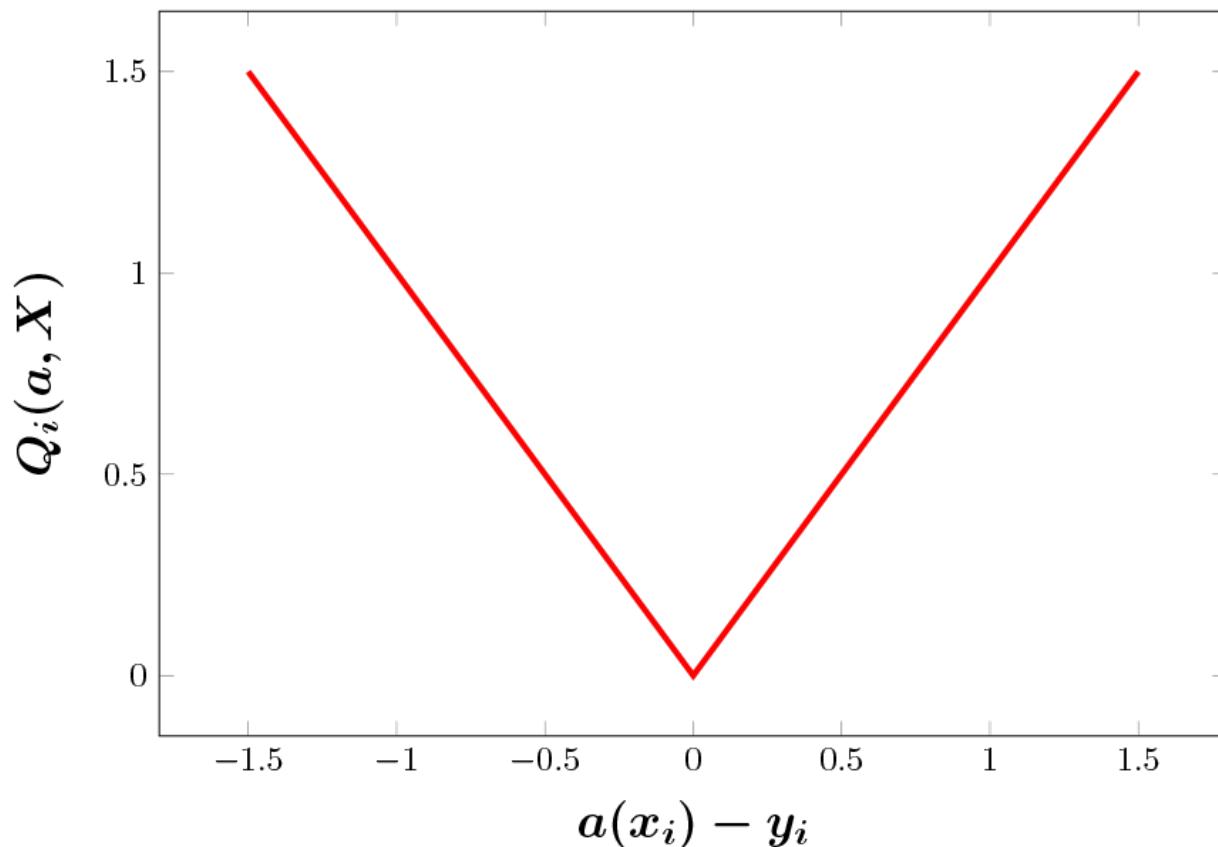
$$Q(\mathbf{w}, \mathbf{X}) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

$$\mathbf{w}_*(x) = \operatorname{argmin}_{\mathbf{w}} Q(\mathbf{w}, \mathbf{X}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

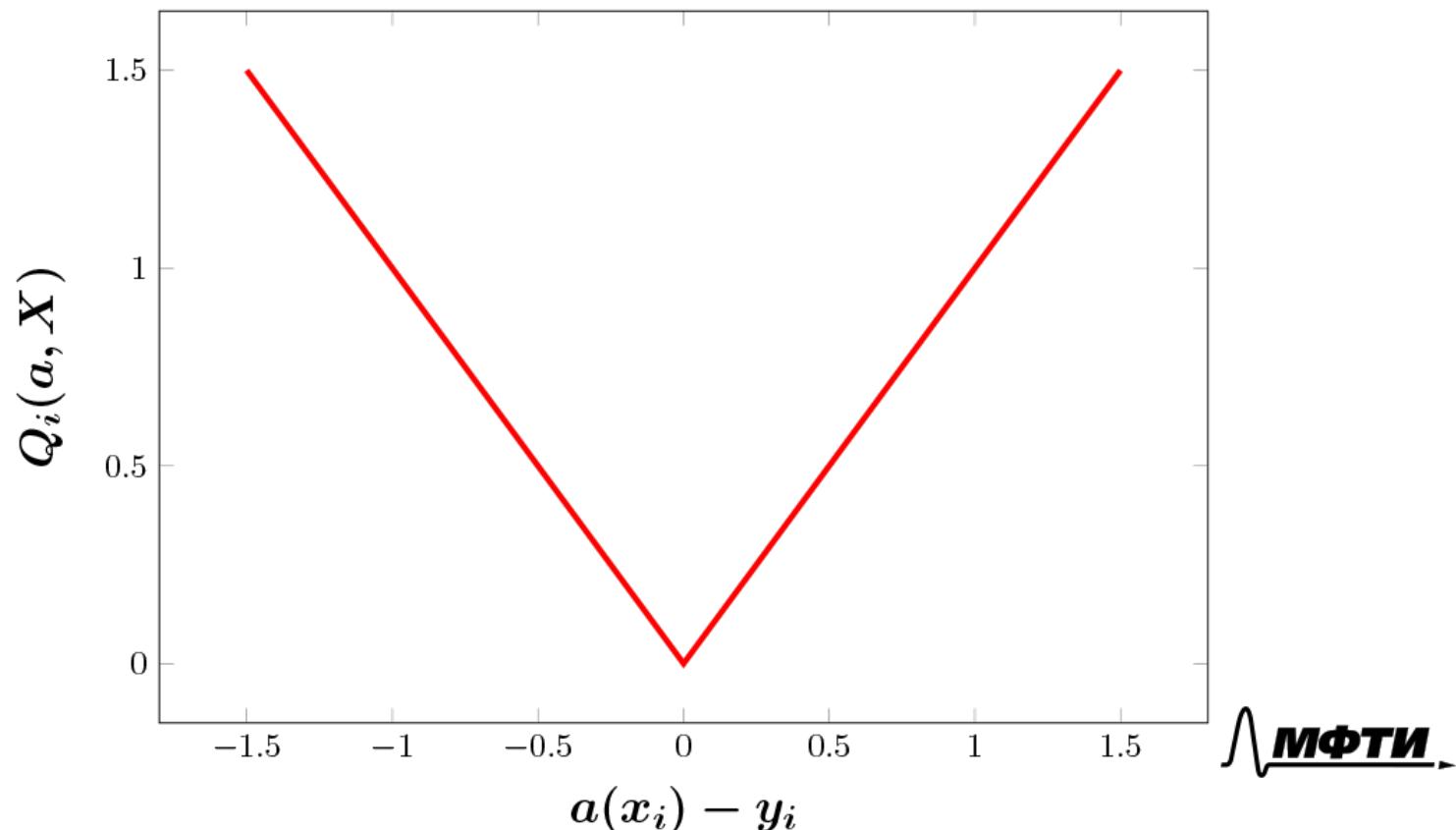
$$a_*(x) = \operatorname{argmin}_a Q(a, X)$$



# СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$
$$a_*(x) = \operatorname{argmin}_a Q(a, X)$$

квантильная  
регрессия



# ДАЛЕЕ В ПРОГРАММЕ

---

- › Метод максимального правдоподобия
- › Свойства регрессии
- › Регуляризация

# МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

---

# КАК ОЦЕНИТЬ НЕИЗВЕСТНЫЙ ПАРАМЕТР ПО ВЫБОРКЕ?

---

$$X \sim F(x, \theta)$$

$$X^n = (X_1, X_2, \dots, X_n)$$

$$\theta — ?$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

---

› (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

# ГИБЕЛЬ КАВАЛЕРИСТОВ

---

› (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$X \sim Pois(\lambda)$$

$$\lambda — ?$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

---

› (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

» (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$P(X = X_i) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

» (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$P(X = X_i) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

$$P(X^n, \lambda) = \prod_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

» (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$P(X = X_i) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

правдоподобие  
выборки

$$P(X^n, \lambda) = \prod_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \equiv L(X^n, \lambda)$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

» (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$P(X = X_i) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

правдоподобие  
выборки

$$P(X^n, \lambda) = \prod_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \equiv L(X^n, \lambda)$$

$$\hat{\lambda}_{\text{ОМП}} = \underset{\lambda}{\operatorname{argmax}} L(X^n, \lambda)$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

» (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$P(X = X_i) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

$$P(X^n, \lambda) = \prod_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \equiv L(X^n, \lambda)$$

оценка  
максимального  
правдоподобия

$$\hat{\lambda}_{\text{ОМП}} = \underset{\lambda}{\operatorname{argmax}} L(X^n, \lambda)$$

# ГИБЕЛЬ КАВАЛЕРИСТОВ

» (Bortkiewicz, 1898):

Количество погибших	0	1	2	3	4	5	Всего
Количество донесений	109	65	22	3	1	0	200

$$P(X = X_i) = \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

$$P(X^n, \lambda) = \prod_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \equiv L(X^n, \lambda)$$

$$\hat{\lambda}_{\text{ОМП}} = \underset{\lambda}{\operatorname{argmax}} L(X^n, \lambda) = \bar{X}_n = 0.61$$

# МЕТОД МАКСИМУМА ПРАВДОПОДОБИЯ

---

$$X \sim F(x, \theta)$$

$$X^n = (X_1, X_2, \dots, X_n)$$

$$L(X^n, \theta) \equiv \prod_{i=1}^n P(X = X_i, \theta)$$

$$\hat{\theta}_{\text{OMP}} = \operatorname{argmax}_{\theta} L(X^n, \theta)$$

# МЕТОД МАКСИМУМА ПРАВДОПОДОБИЯ

---

$$X \sim F(x, \theta)$$

$$X^n = (X_1, X_2, \dots, X_n)$$

$$\ln L(X^n, \theta) = \sum_{i=1}^n \ln P(X = X_i, \theta)$$

$$\hat{\theta}_{\text{ОМП}} = \operatorname{argmax}_{\theta} \ln L(X^n, \theta)$$

# ДЛЯ НЕПРЕРЫВНОГО РАСПРЕДЕЛЕНИЯ

---

$$X \sim f(x, \theta)$$

$$X^n = (X_1, X_2, \dots, X_n)$$

$$L(X^n, \theta) \equiv \prod_{i=1}^n f(X_i, \theta)$$

$$\hat{\theta}_{\text{ОМП}} = \operatorname{argmax}_{\theta} L(X^n, \theta)$$

# ПОЛЕЗНЫЕ СВОЙСТВА ОМП

---

› Состоятельность:

при  $n \rightarrow \infty$   $\hat{\theta}_{\text{ОМП}} \rightarrow \theta$

› Асимптотическая нормальность:

при  $n \rightarrow \infty$   $\hat{\theta}_{\text{ОМП}} \sim N(\theta, I^{-1}(\theta))$

# РЕЗЮМЕ

---

- › Максимизация правдоподобия — полезный метод оценки неизвестных параметров распределений

# ДАЛЕЕ В ПРОГРАММЕ

---

› При чём тут регрессия?

# РЕГРЕССИЯ КАК МАКСИМИЗАЦИЯ ПРАВДОПОДОБИЯ

---

# ЧТО МЫ ПОЛУЧАЕМ, МИНИМИЗИРУЯ СРЕДНЕКВАДРАТИЧНУЮ ОШИБКУ?

---

» МНК-регрессия:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$a_*(x) = \operatorname{argmin}_a Q(a, X)$$

# МОДЕЛЬ ШУМА

---

$$y \approx a(x)$$

# МОДЕЛЬ ШУМА

---

$$y = a(x) + \varepsilon,$$

$\varepsilon$  — случайный шум.

# МОДЕЛЬ ШУМА

---

$$y = a(x) + \varepsilon,$$

$\varepsilon$  — случайный шум.

Пусть  $\varepsilon \sim N(0, \sigma^2)$

# МОДЕЛЬ ШУМА

---

$$y = a(x) + \varepsilon,$$

$\varepsilon$  — случайный шум.

Пусть  $\varepsilon \sim N(0, \sigma^2) \Rightarrow$

$$a_*(x) = \operatorname{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 —$$

оценка максимального правдоподобия  
для  $a(x)$

# И ЧТО?

---

› Благодаря свойствам ОМП мы можем:

# И ЧТО?

---

Благодаря свойствам ОМП мы можем:

- ▶ определять значимость признаков  $x^j$  и делать их отбор

# И ЧТО?

---

Благодаря свойствам ОМП мы можем:

- ▶ определять значимость признаков  $x^j$  и делать их отбор
- ▶ строить доверительные интервалы для значения отклика на новых объектах

# МОДЕЛЬ ШУМА

---

$$y = a(x) + \varepsilon,$$

$\varepsilon$  — случайный шум.

- › Пусть  $\varepsilon$  имеет распределение Лапласа с нулевым средним:

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x|}$$

# МОДЕЛЬ ШУМА

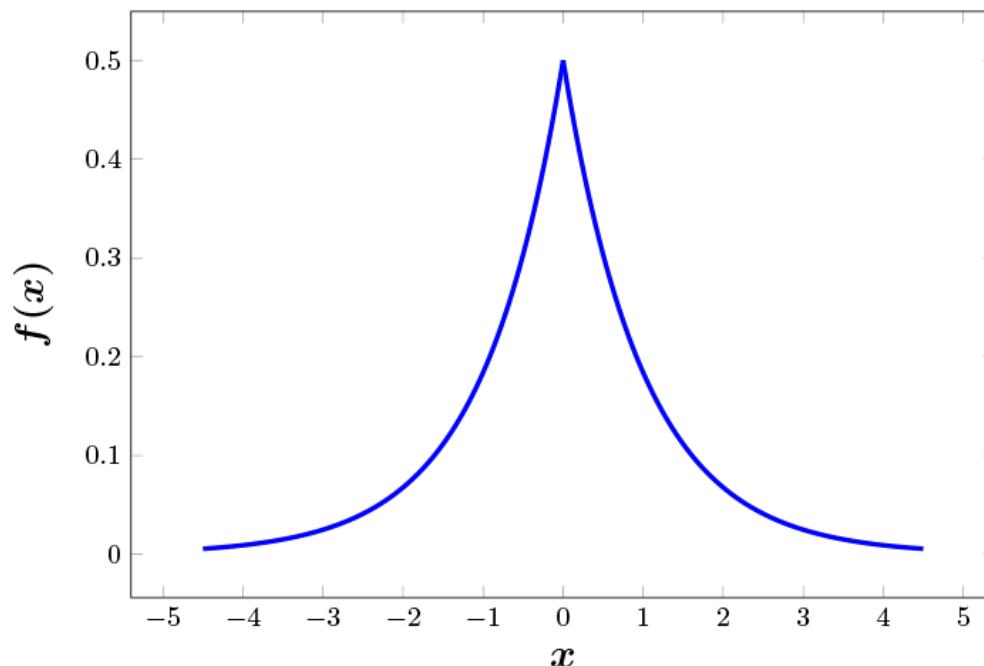
---

$$y = a(x) + \varepsilon,$$

$\varepsilon$  — случайный шум.

- › Пусть  $\varepsilon$  имеет распределение Лапласа с нулевым средним:

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x|}$$



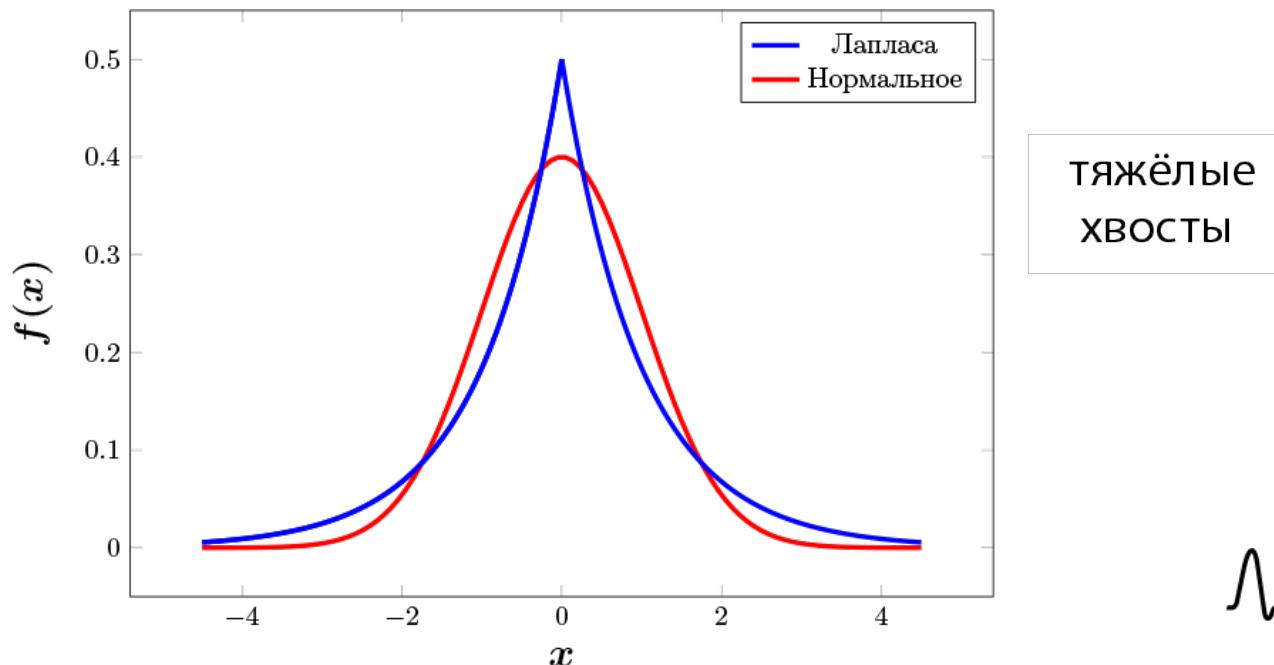
# МОДЕЛЬ ШУМА

$$y = a(x) + \varepsilon,$$

$\varepsilon$  — случайный шум.

- › Пусть  $\varepsilon$  имеет распределение Лапласа с нулевым средним:

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x|}$$



# МОДЕЛЬ ШУМА

---

$$y = a(x) + \varepsilon,$$

$\varepsilon$  — случайный шум.

- › Пусть  $\varepsilon$  имеет распределение Лапласа с нулевым средним  $\Rightarrow$

$$a_*(x) = \operatorname{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

оценка максимального правдоподобия для  $a(x)$

# РЕЗЮМЕ

---

- › МНК-регрессия даёт ОМП для  $a(x)$ , если шум нормальный
- › Регрессия со средней абсолютной ошибкой даёт ОМП для  $a(x)$ , если шум лапласовский

# ДАЛЕЕ В ПРОГРАММЕ

---

- › Регрессия как оценка среднего

# РЕГРЕССИЯ КАК ОЦЕНКА СРЕДНЕГО

---

# МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

---

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$a_*(x) = \operatorname*{argmin}_a Q(a, X)$$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a - y_i)^2$$

› Пусть  $a$  — константа

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a) = \int_t (a - t)^2 f(t) dt$$

- » Пусть  $a$  — константа
- » Пусть  $\ell = \infty$ , то есть, у нас не выборка из  $y$ , а вся случайная величина  $y \sim f(t)$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a) = \int_t (a - t)^2 f(t) dt$$

- » Пусть  $a$  — константа
- » Пусть  $\ell = \infty$ , то есть, у нас не выборка из  $y$ , а вся случайная величина  $y \sim f(t)$

$$\Rightarrow a_* = \operatorname{argmin}_a Q(a) = \mathbb{E}y$$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a(x), X) = \int_t (a(x) - t)^2 f(t) dt$$

- › Пусть  $a$  — произвольная функция от  $x$
- › Пусть  $\ell = \infty$ , то есть, у нас не выборка из  $y$ , а вся случайная величина  $y \sim f(t)$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a(x), X) = \int_t (a(x) - t)^2 f(t) dt$$

- » Пусть  $a$  — произвольная функция от  $x$
- » Пусть  $\ell = \infty$ , то есть, у нас не выборка из  $y$ , а вся случайная величина  $y \sim f(t)$

$$\Rightarrow a_* = \operatorname{argmin}_a Q(a) = \mathbb{E}(y|x)$$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a(x), X) = \int_t (a(x) - t)^2 f(t) dt$$

- » Пусть  $a$  — произвольная функция от  $x$
- » Пусть  $\ell = \infty$ , то есть, у нас не выборка из  $y$ , а вся случайная величина  $y \sim f(t)$

$$\Rightarrow a_* = \operatorname{argmin}_a Q(a) = \mathbb{E}(y|x)$$

средний  $y$  при таком  $x$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(a(x), X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$a_*(x) = \operatorname{argmin} Q(a, X)$  — лучшая  
аппроксимация  $\mathbb{E}(y|x)$

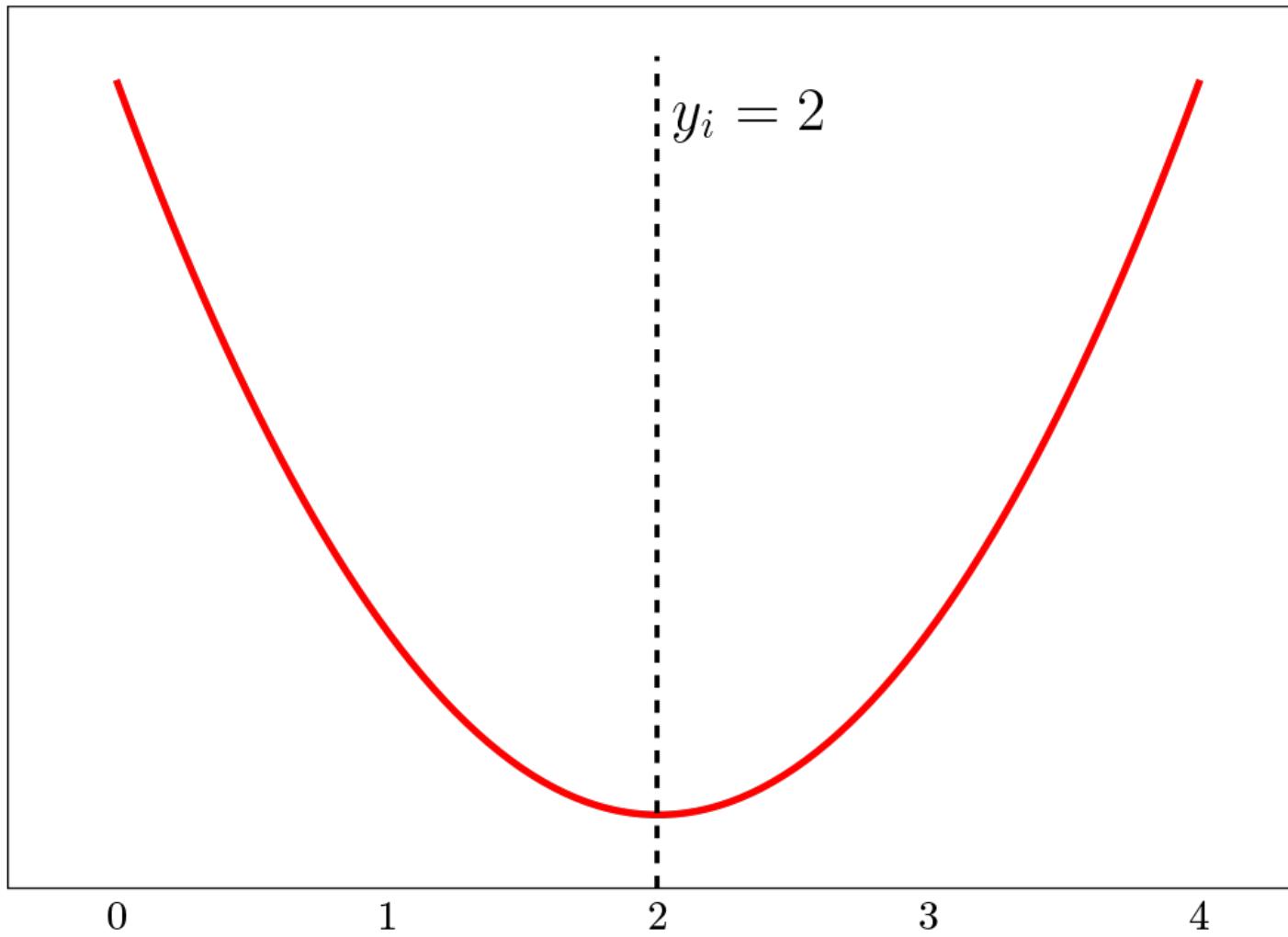
# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА

---

$$Q(\mathbf{w}, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

$\langle \mathbf{w}_*, \mathbf{x}_i \rangle$ , где  $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}} Q(\mathbf{w}, X)$  —  
лучшая линейная аппроксимация  $\mathbb{E}(y|\mathbf{x})$

# СРЕДНЕКВАДРАТИЧНАЯ ОШИБКА



$a(x_i)$

# ДИВЕРГЕНЦИЯ БРЕГМАНА

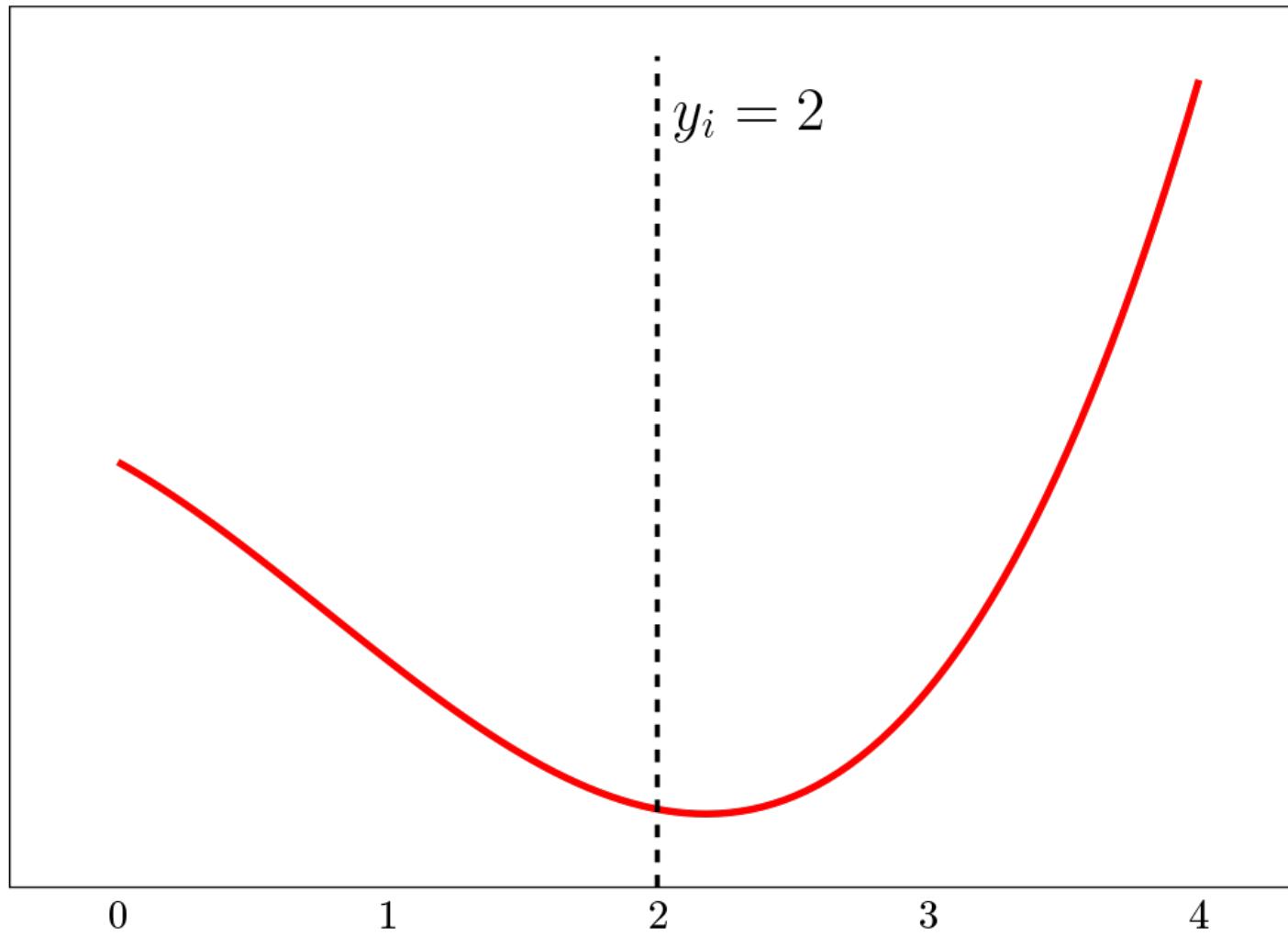
---

$$Q(a, X) = \\ = \varphi(y) - \varphi(a(X)) - \varphi'(a(X))(y - a(X))$$

$\varphi$  — любая непрерывно дифференцируемая выпуклая функция.

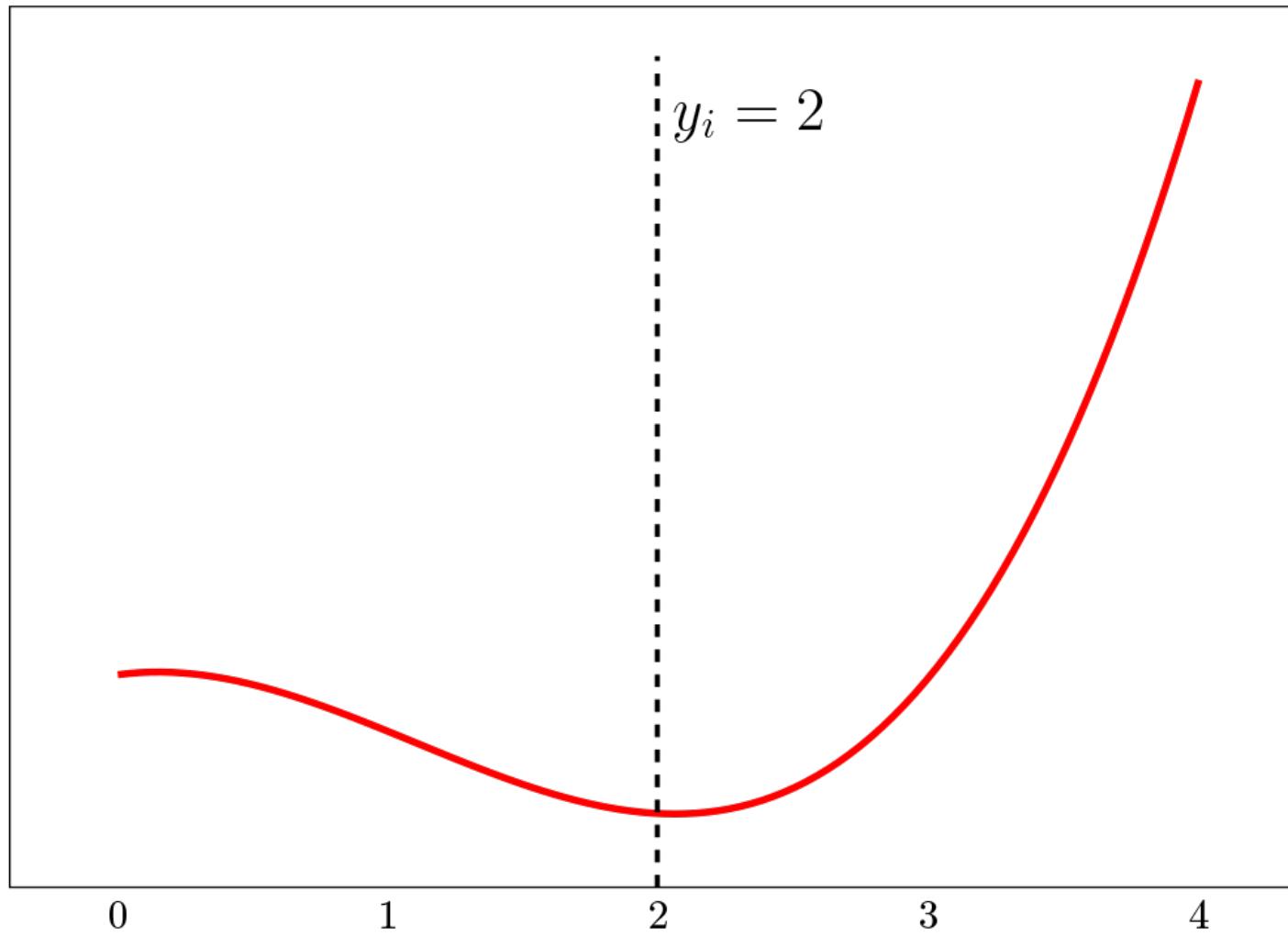
$a_*(x) = \operatorname{argmin}_a Q(a, X)$  — лучшая аппроксимация  $\mathbb{E}(y|x)$

# ДИВЕРГЕНЦИЯ БРЕГМАНА



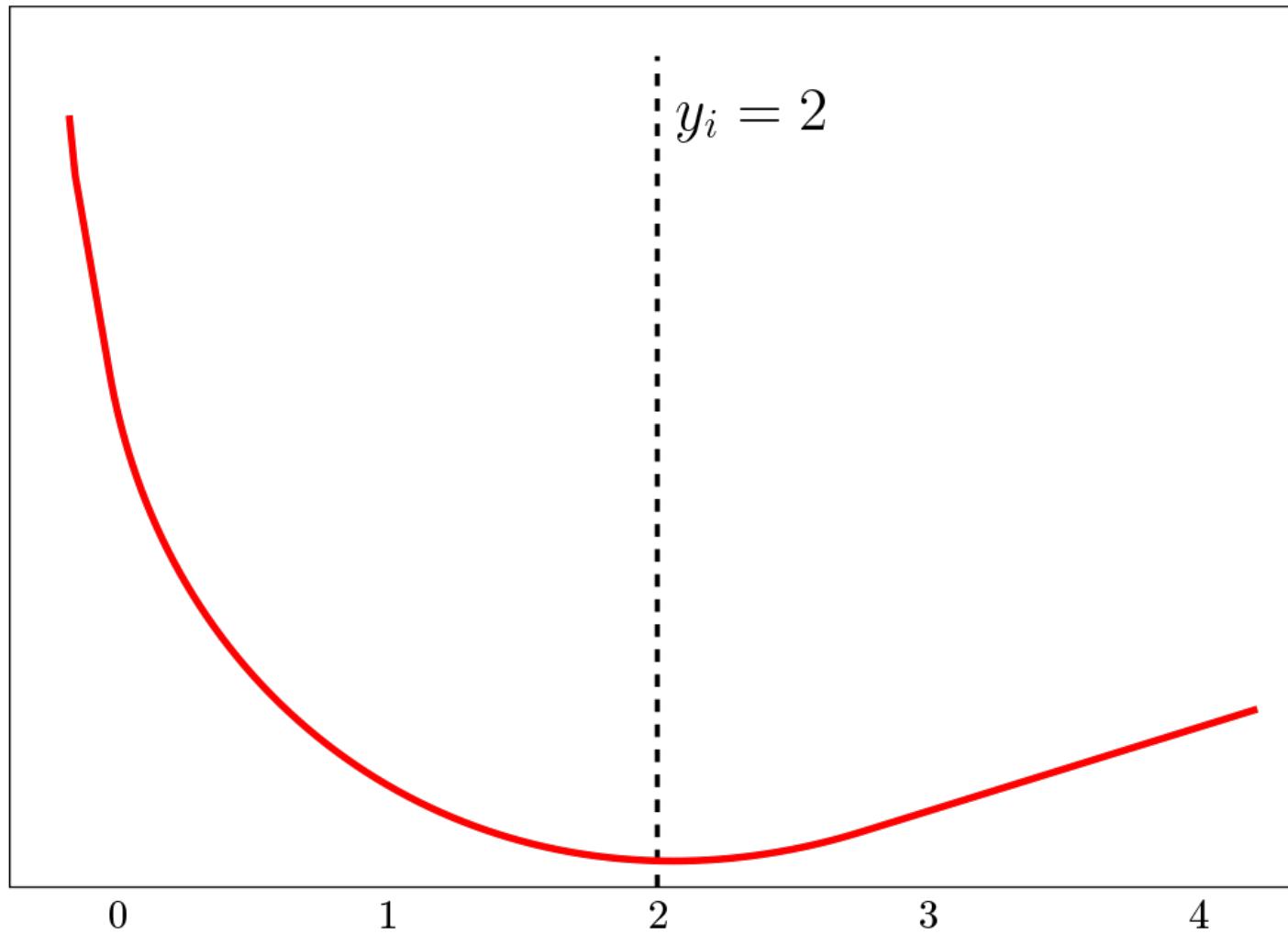
$a(x_i)$

# ДИВЕРГЕНЦИЯ БРЕГМАНА



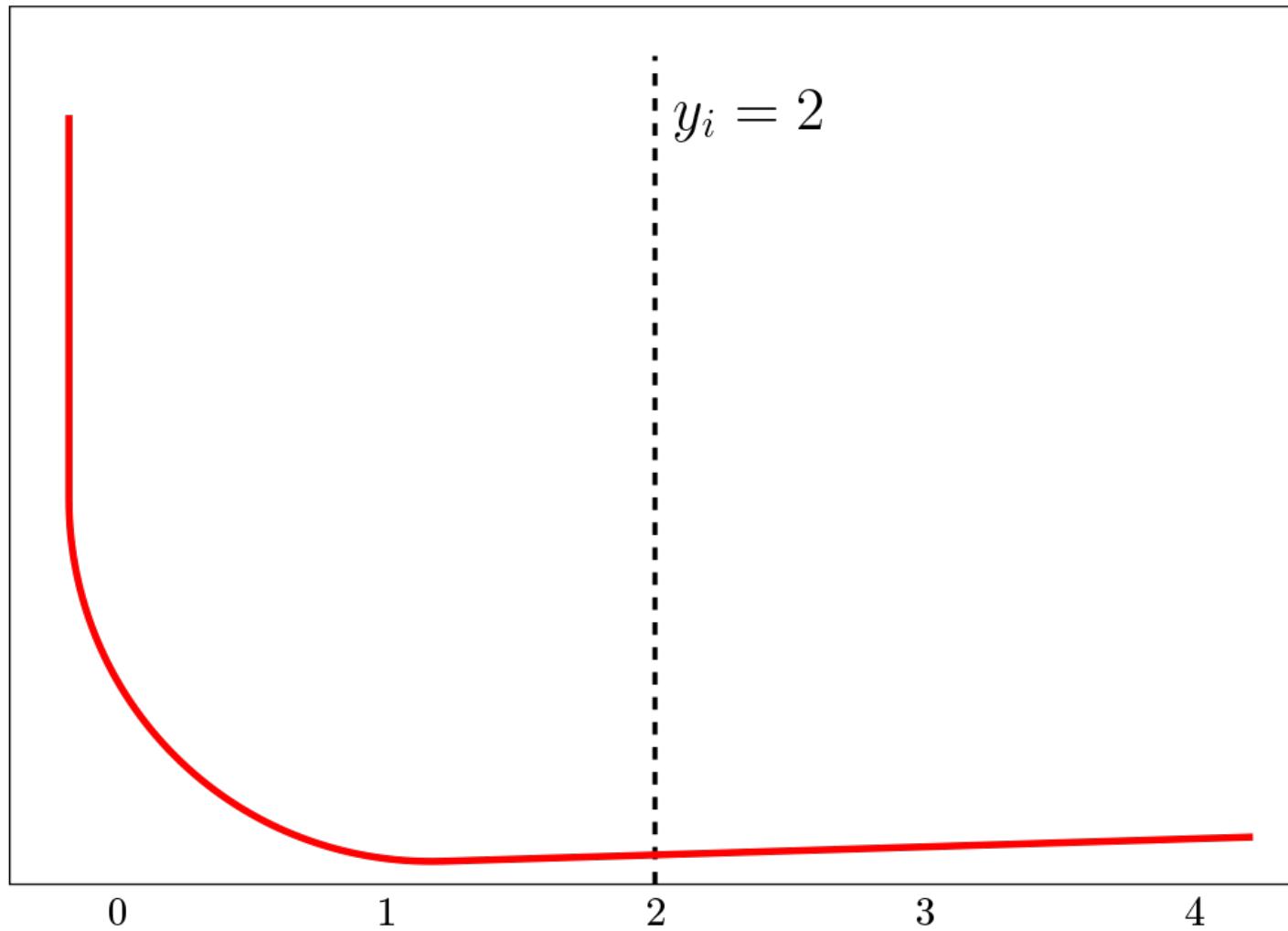
$a(x_i)$

# ДИВЕРГЕНЦИЯ БРЕГМАНА



$a(x_i)$

# ДИВЕРГЕНЦИЯ БРЕГМАНА

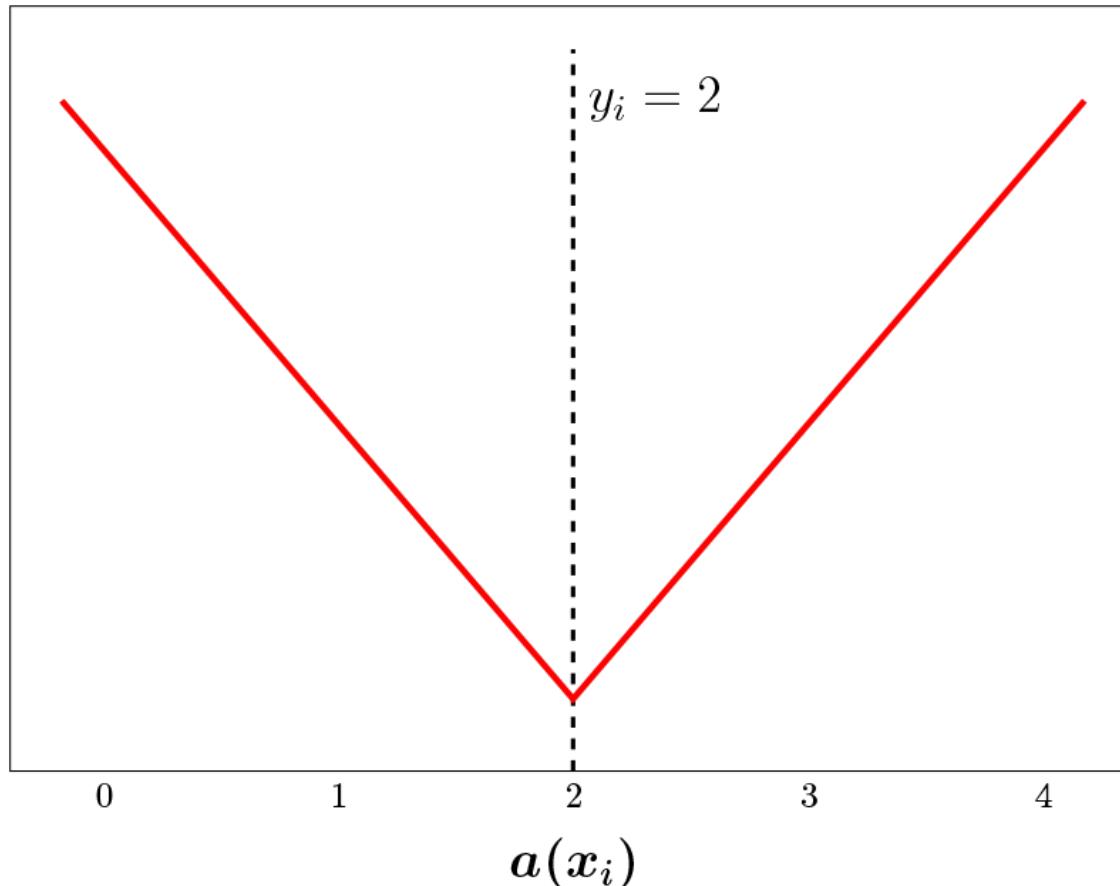


$a(x_i)$

# СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА

---

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$



# СРЕДНЯЯ АБСОЛЮТНАЯ ОШИБКА

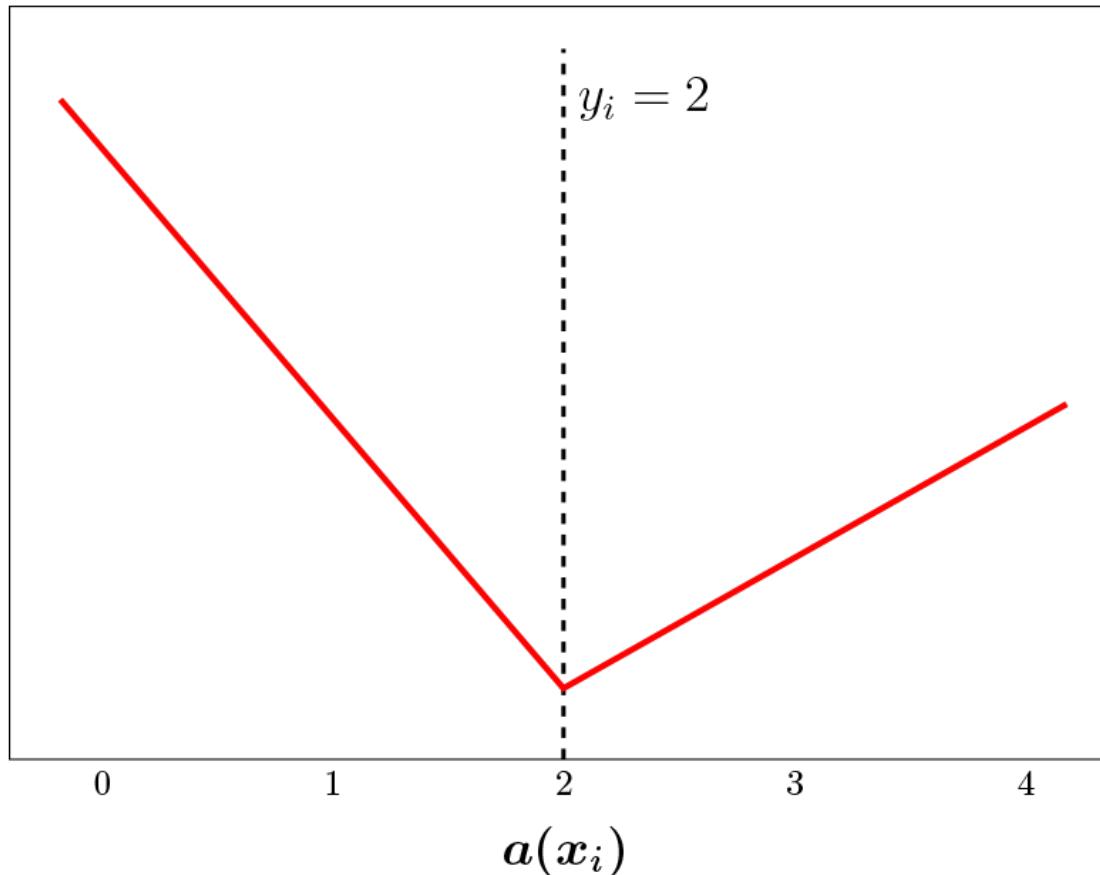
---

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

$a_*(x) = \operatorname{argmin}_a Q(a, X)$  — лучшая  
аппроксимация  $\operatorname{med}(y|x)$

# НЕСИММЕТРИЧНАЯ АБСОЛЮТНАЯ ОШИБКА

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} ((\tau - 1)[y_i < a(x_i)] + \\ + \tau[y_i \geq a(x_i)])(y_i - a(x_i))$$



# НЕСИММЕТРИЧНАЯ АБСОЛЮТНАЯ ОШИБКА

---

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} ((\tau - 1)[y_i < a(x_i)] + \\ + \tau[y_i \geq a(x_i)])(y_i - a(x_i))$$

$a_*(x) = \operatorname{argmin}_a Q(a, X)$  — лучшая  
аппроксимация  $y|x$  порядка  $\tau$

# РЕЗЮМЕ

---

- › Решение задачи МНК-регрессии — оценка условного матожидания  $\mathbb{E}(y|x)$
- › Решение задачи квантильной регрессии — оценка условного квантиля  $y|x$ ; при использовании средней абсолютной ошибки — условной медианы  $\text{med}(y|x)$

# ДАЛЕЕ В ПРОГРАММЕ

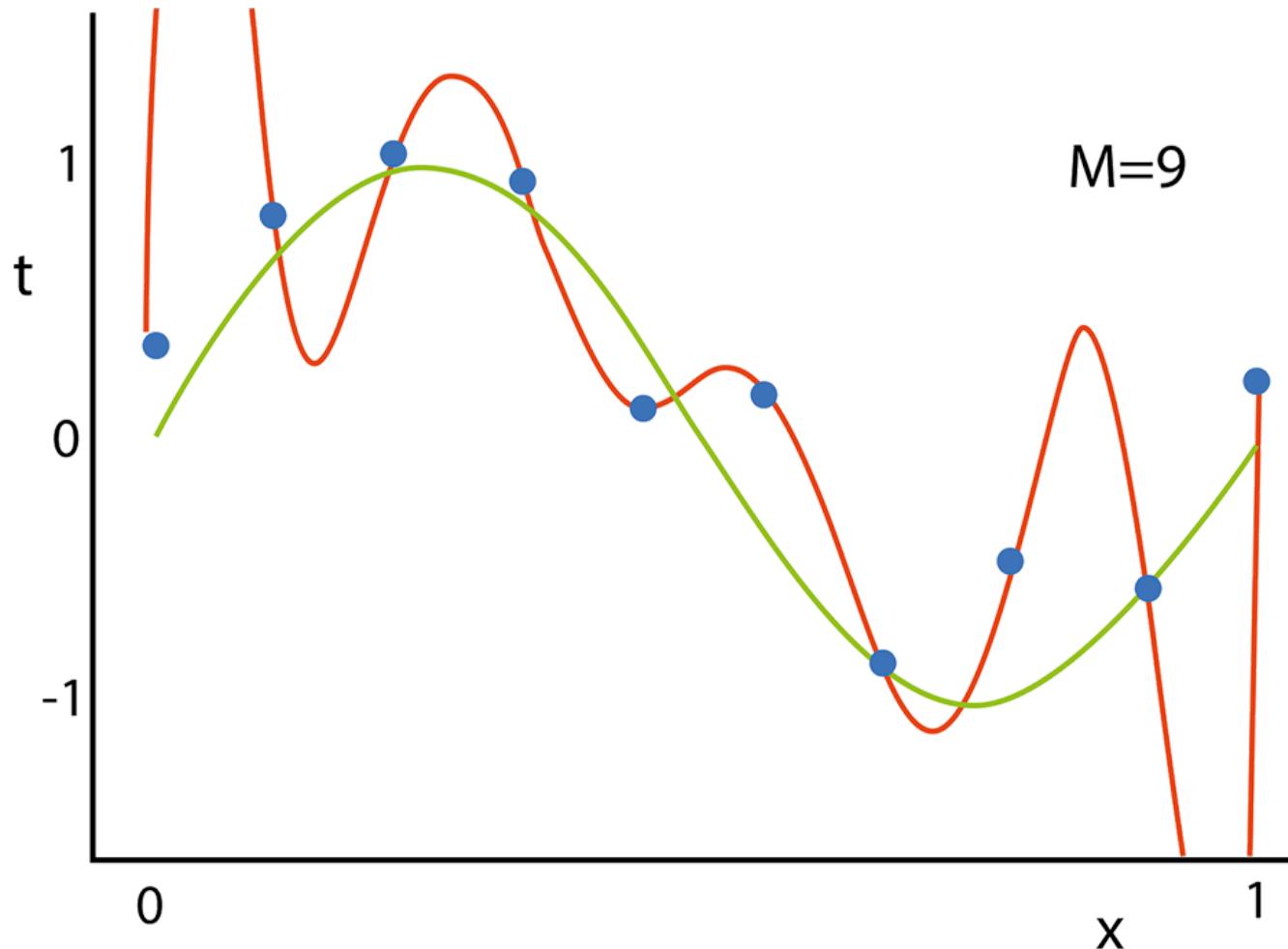
---

› Регуляризация

# РЕГУЛЯРИЗАЦИЯ

---

# ПЕРЕОБУЧЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ



# ПЕРЕОБУЧЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ

---

Решения:

- ▶ больше данных

# ПЕРЕОБУЧЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ

---

Решения:

- ▶ больше данных
- ▶ меньше признаков

# ПЕРЕОБУЧЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ

---

Решения:

- ▶ больше данных
- ▶ меньше признаков
- ▶ ограничить веса

# ОГРАНИЧИТЬ ВЕСА

---

»  $L_2$ -регуляризатор

$$\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d \mathbf{w}_j^2 \right)$$

»  $L_1$ -регуляризатор

$$\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d |\mathbf{w}_j| \right)$$

# ОГРАНИЧИТЬ ВЕСА

---

»  $L_2$ -регуляризатор

гребневая регрессия

$$\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d \mathbf{w}_j^2 \right)$$

»  $L_1$ -регуляризатор

лассо

$$\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d |\mathbf{w}_j| \right)$$

# ОГРАНИЧИТЬ ВЕСА

---

»  $L_2$ -регуляризатор

$$\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d \mathbf{w}_j^2 \right)$$

»  $L_1$ -регуляризатор

$$\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w}} \left( \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 + \lambda \sum_{j=1}^d |\mathbf{w}_j| \right)$$

Важно: константное слагаемое не входит в штрафы

# МОДЕЛЬНЫЙ ПРИМЕР

---

- › Пусть  $\ell = d$ ,  $X$  — единичная матрица, константы нет.
- › МНК-регрессия:

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{\ell} (\mathbf{w}_i - y_i)^2$$

$$\mathbf{w}_{*j} = y_j$$

# МОДЕЛЬНЫЙ ПРИМЕР

---

- › Пусть  $\ell = d$ ,  $X$  — единичная матрица, константы нет.
- › МНК-регрессия:

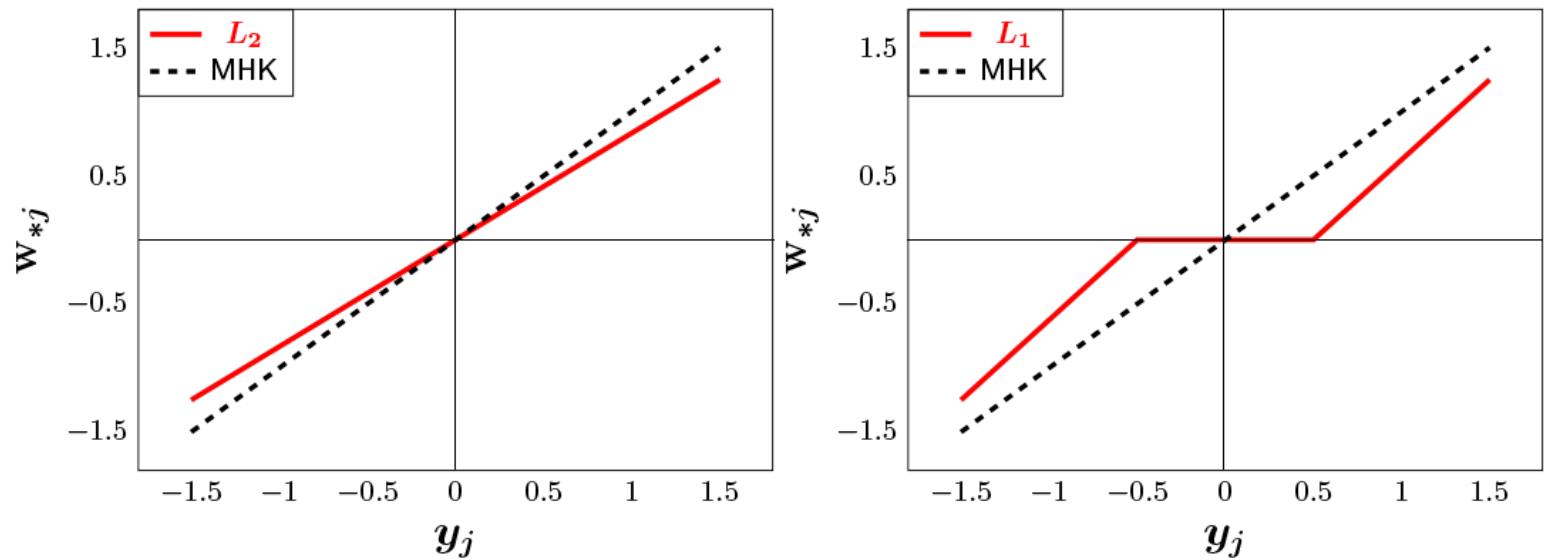
$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{\ell} (\mathbf{w}_i - y_i)^2$$

- › Гребневая регрессия:  $\mathbf{w}_{*j} = \frac{y_j}{1+\lambda}$

- › Лассо:  $\mathbf{w}_{*j} = \begin{cases} y_j - \lambda/2, & y_j > \lambda/2 \\ y_j + \lambda/2, & y_j < -\lambda/2 \\ 0, & |y_j| \leq \lambda/2 \end{cases}$

# МОДЕЛЬНЫЙ ПРИМЕР

› Пусть  $\ell = d$ ,  $X$  — единичная матрица, константы нет.



# СМЕЩЕНИЕ И ДИСПЕРСИЯ (BIAS-VARIANCE TRADEOFF)

---

- › Матожидание квадрата ошибки регрессии:  
 $\mathbb{E}(a_*(x) - y)^2$

# СМЕЩЕНИЕ И ДИСПЕРСИЯ (BIAS-VARIANCE TRADEOFF)

---

› Матожидание квадрата ошибки регрессии:

$$\begin{aligned}\mathbb{E}(a_*(x) - y)^2 &= \\ &= (\mathbb{E}a_*(x) - a(x))^2 + \mathbb{D}a_*(x) + \sigma^2\end{aligned}$$

# СМЕЩЕНИЕ И ДИСПЕРСИЯ (BIAS-VARIANCE TRADEOFF)

---

› Матожидание квадрата ошибки регрессии:

$$\mathbb{E}(a_*(x) - y)^2 =$$

$$= (\mathbb{E}a_*(x) - a(x))^2 + \mathbb{D}a_*(x) + \sigma^2$$

смещение

дисперсия

шум

# СМЕЩЕНИЕ И ДИСПЕРСИЯ (BIAS-VARIANCE TRADEOFF)

---

- » Матожидание квадрата ошибки регрессии:

$$\mathbb{E}(a_*(x) - y)^2 = \\ = (\mathbb{E}a_*(x) - a(x))^2 + \mathbb{D}a_*(x) + \sigma^2$$

смещение

дисперсия

шум

- » МНК-оценки имеют нулевое смещение

# СМЕЩЕНИЕ И ДИСПЕРСИЯ (BIAS-VARIANCE TRADEOFF)

---

- » Матожидание квадрата ошибки регрессии:

$$\mathbb{E}(a_*(x) - y)^2 = \\ = (\mathbb{E}a_*(x) - a(x))^2 + \mathbb{D}a_*(x) + \sigma^2$$

смещение

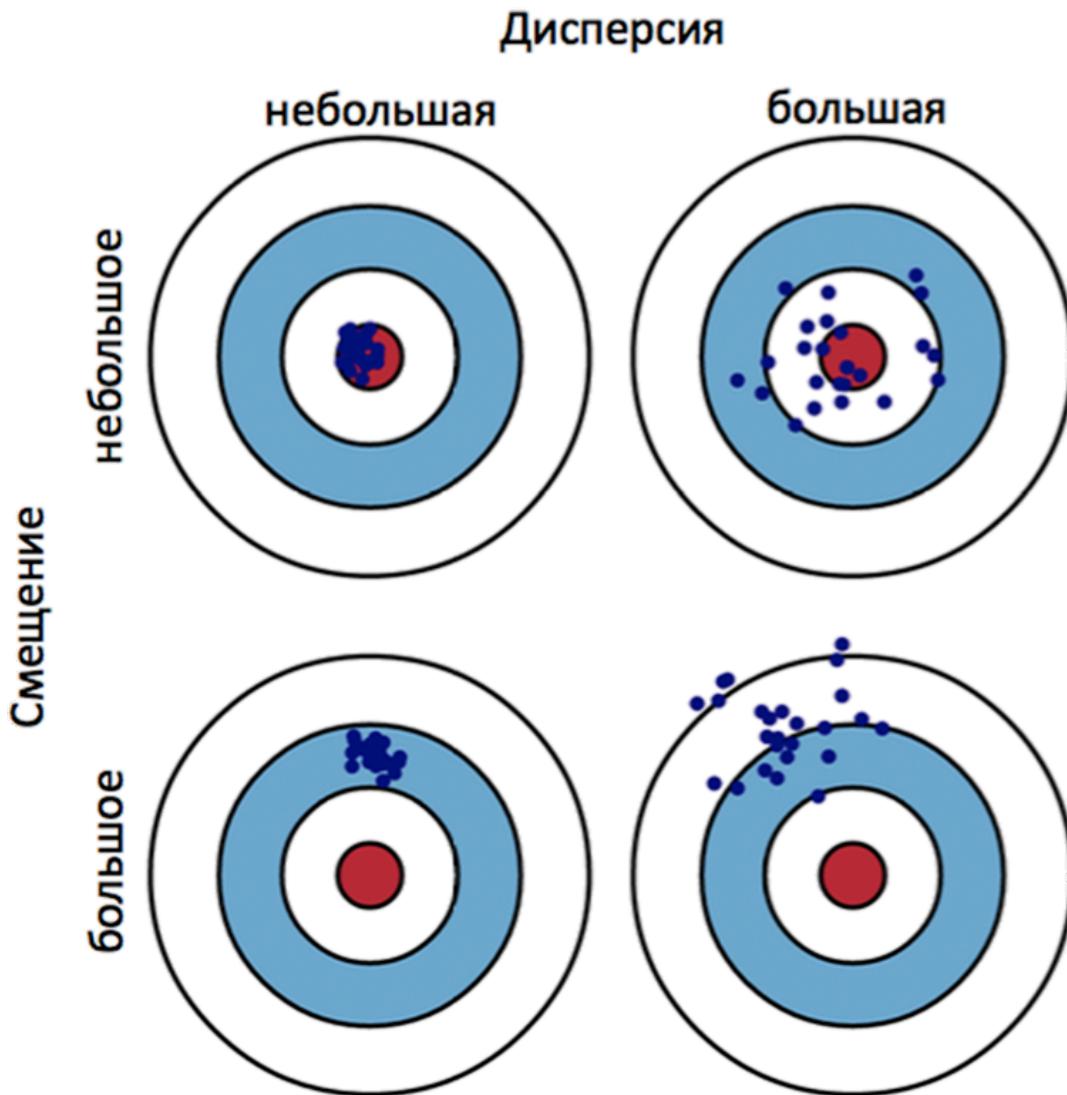
дисперсия

шум

- » МНК-оценки имеют нулевое смещение

- » Регуляризация даёт смещённые оценки, но их дисперсия может быть меньше!

# СМЕЩЕНИЕ И ДИСПЕРСИЯ (BIAS-VARIANCE TRADEOFF)



# В БАЙЕСОВСКОЙ ПОСТАНОВКЕ

---

- › Гребневая регрессия соответствует заданию нормального априорного распределения на коэффициенты
- › Лассо соответствует заданию Лапласовского априорного распределения на коэффициенты  
(подробности скоро!)

# РЕШЕНИЕ

---

- › Задача гребневой регрессии имеет аналитическое решение:

$$\mathbf{w}_* = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

- › Для решения лассо существует эффективный численный метод

# РЕЗЮМЕ

---

- › Регуляризация — один из способов борьбы с переобучением
- › Даёт смещённые оценки коэффициентов, но ошибка может быть меньше за счёт меньшей дисперсии
- › Лассо ещё и отбирает признаки

# ДАЛЕЕ В ПРОГРАММЕ

---

› Логистическая регрессия

# ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

---

# БИНАРНАЯ КЛАССИФИКАЦИЯ

---

- »  $\mathbb{X}$  — пространство объектов,  
 $\mathbb{Y}$  — пространство ответов
- »  $x = (x^1, \dots, x^d)$  — признаковое описание
- »  $X = (x_i, y_i)_{i=1}^\ell$  — обучающая выборка
- »  $\mathbb{Y} = \{0, 1\}$

# ЛИНЕЙНАЯ РЕГРЕССИЯ

---

$$Q(\mathbf{w}, \mathbf{X}) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2$$

$$\mathbf{w}_* = \operatorname*{argmin}_{\mathbf{w}} Q(\mathbf{w}, \mathbf{X})$$

- » Предсказываем  $y = 1$ , если  $\langle \mathbf{w}, \mathbf{x} \rangle > 0.5$
- » Это аналог линейного дискриминанта Фишера

# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

---

- » Хотим оценить  $P(y = 1|x) \equiv \pi(x)$

# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

---

- › Хотим оценить  $P(y = 1|x) \equiv \pi(x)$
- › Линейная регрессия:  $\pi(x) \approx \langle w, x \rangle$

# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

---

- » Хотим оценить  $P(y = 1|x) \equiv \pi(x)$
- » Линейная регрессия:  $\pi(x) \approx \langle w, x \rangle$ 
  - $\pi(x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x) = \mathbb{E}(y|x)$

# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

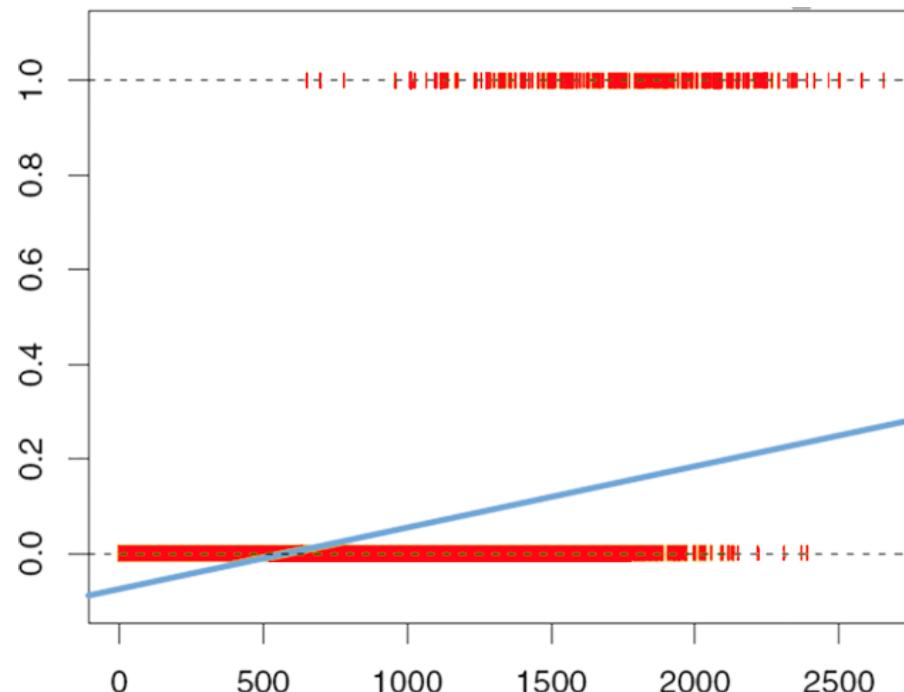
---

- » Хотим оценить  $P(y = 1|x) \equiv \pi(x)$
- » Линейная регрессия:  $\pi(x) \approx \langle w, x \rangle$ 
  - $\pi(x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x) = \mathbb{E}(y|x)$
  - $\langle w, x \rangle$  не обязательно лежит в  $[0,1]$

# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

---

- » Хотим оценить  $P(y = 1|x) \equiv \pi(x)$
- » Линейная регрессия:  $\pi(x) \approx \langle w, x \rangle$ 
  - $\pi(x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x) = \mathbb{E}(y|x)$
  - $\langle w, x \rangle$  не обязательно лежит в  $[0,1]$



# ИДЕЯ

---

- › Возьмём какую-то функцию  $g$ , которая переводит  $[0,1]$  в  $\mathbb{R}$  и построим линейную регрессию

$$g(\mathbb{E}(y|x)) \approx \langle \mathbf{w}, x \rangle$$

$$\mathbb{E}(y|x) \approx g^{-1}(\langle \mathbf{w}, x \rangle)$$

# ИДЕЯ

---

- › Возьмём какую-то функцию  $g$ , которая переводит  $[0,1]$  в  $\mathbb{R}$  и построим линейную регрессию

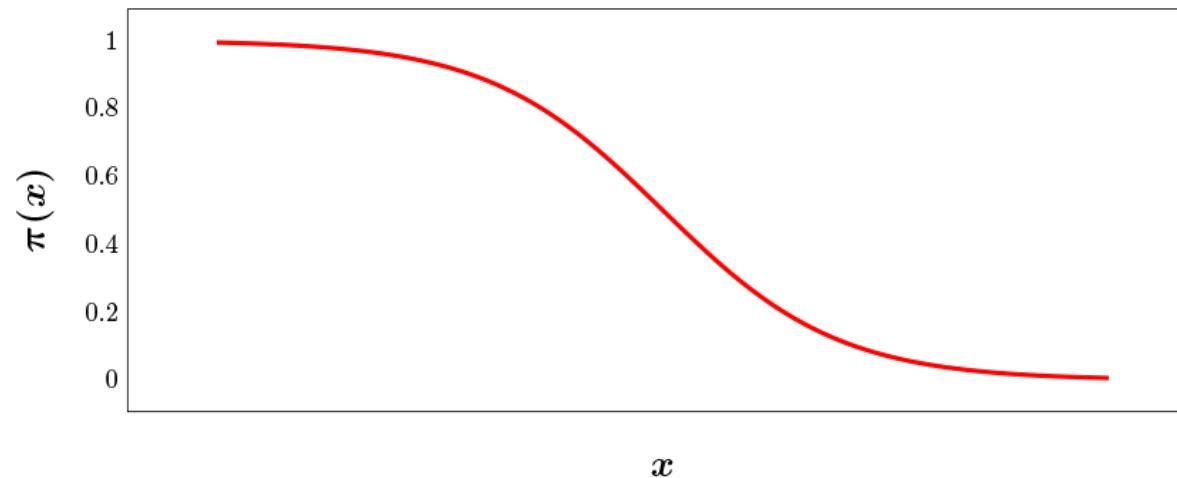
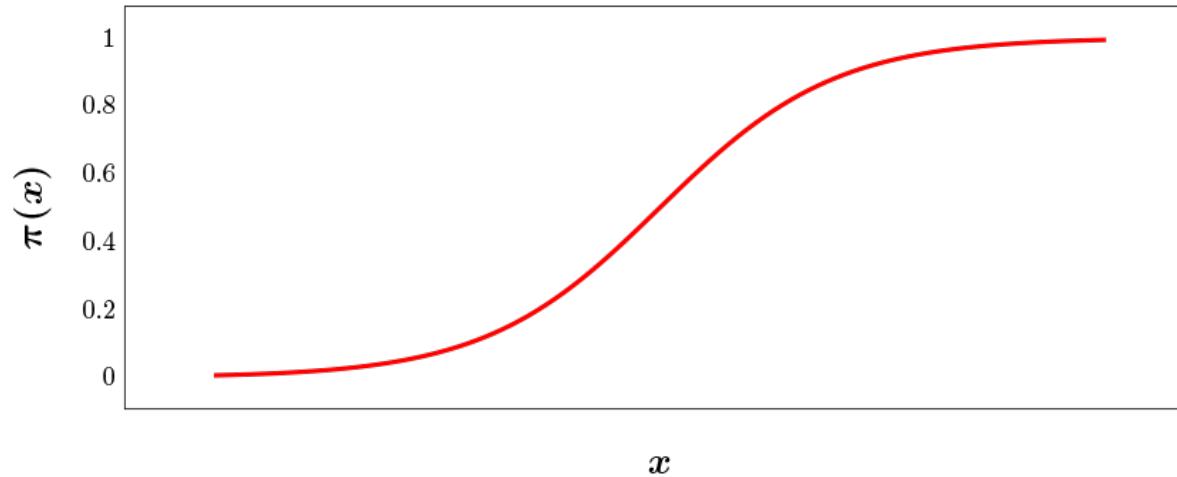
$$g(\mathbb{E}(y|x)) \approx \langle \mathbf{w}, x \rangle$$

$$\mathbb{E}(y|x) \approx g^{-1}(\langle \mathbf{w}, x \rangle)$$

обобщённая линейная  
модель (GLM)

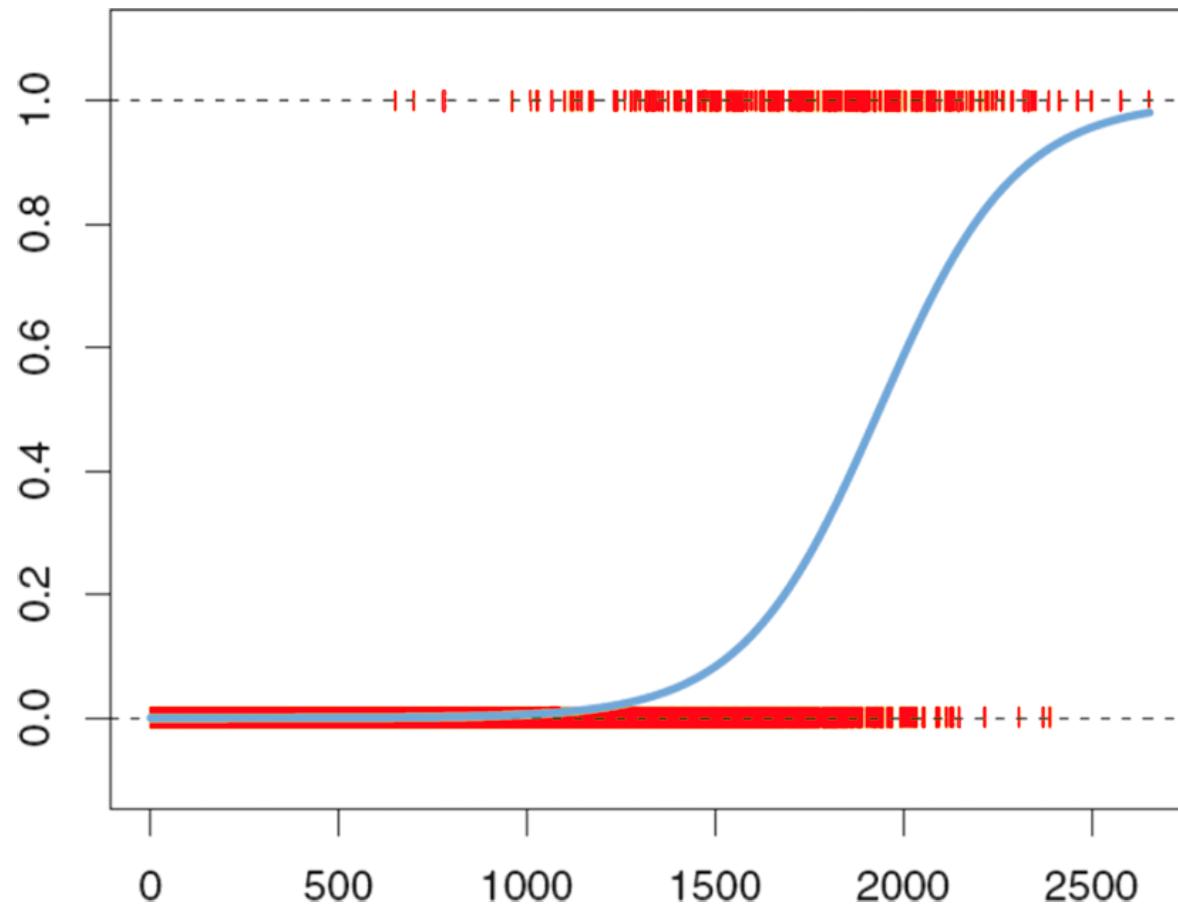
# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

» Логистическая регрессия:  $\pi(x) \approx \frac{e^{\langle w, x \rangle}}{1 + e^{\langle w, x \rangle}}$



# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

› Логистическая регрессия:  $\pi(x) \approx \frac{e^{\langle w, x \rangle}}{1 + e^{\langle w, x \rangle}}$



# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

---

› Логистическая регрессия:

$$\pi(x) \approx \frac{e^{\langle w, x \rangle}}{1 + e^{\langle w, x \rangle}} \Leftrightarrow \ln \frac{\pi(x)}{1 - \pi(x)} \approx \langle w, x \rangle$$

# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

---

› Логистическая регрессия:

$$\pi(x) \approx \frac{e^{\langle w, x \rangle}}{1 + e^{\langle w, x \rangle}} \Leftrightarrow \ln \frac{\pi(x)}{1 - \pi(x)} \approx \langle w, x \rangle$$

риск

# ПРЕДСКАЗАНИЕ ВЕРОЯТНОСТЕЙ

---

› Логистическая регрессия:

$$\pi(x) \approx \frac{e^{\langle w, x \rangle}}{1 + e^{\langle w, x \rangle}} \Leftrightarrow \ln \frac{\pi(x)}{1 - \pi(x)} \approx \langle w, x \rangle$$

ЛОГИТ

# ОЦЕНКА ПАРАМЕТРОВ

---

› Правдоподобие обучающей выборки:

$$L(X) = \prod_{i:y_i=1} \pi(x_i) \prod_{i:y_i=0} (1 - \pi(x_i))$$

$$\begin{aligned}\ln L(X) &= \sum_{i=1}^{\ell} \ln \left( \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \right) = \\ &= \sum_{i=1}^{\ell} \left( y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right)\end{aligned}$$

# ОЦЕНКА ПАРАМЕТРОВ

---

$$-\ln L(X) = - \sum_{i=1}^{\ell} \left( y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right)$$

Log-loss, кросс-энтропия

# ОЦЕНКА ПАРАМЕТРОВ

---

$$-\ln L(\mathbf{X}) = -\sum_{i=1}^{\ell} \left( y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right)$$

Если переобозначить  $y = 0$  за  $-1$ , то путём несложных преобразований можно получить логистическую функцию потерь

$$Q(\mathbf{w}, \mathbf{X}) = \sum_{i=1}^{\ell} \ln (1 + \exp (-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle))$$

# ОЦЕНКА ПАРАМЕТРОВ

---

$$-\ln L(\mathbf{X}) = -\sum_{i=1}^{\ell} \left( y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right)$$

Если переобозначить  $y = 0$  за  $-1$ , то путём несложных преобразований можно получить логистическую функцию потерь

$$Q(\mathbf{w}, \mathbf{X}) = \sum_{i=1}^{\ell} \ln (1 + \exp (-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle))$$

# ОЦЕНКА ПАРАМЕТРОВ

---

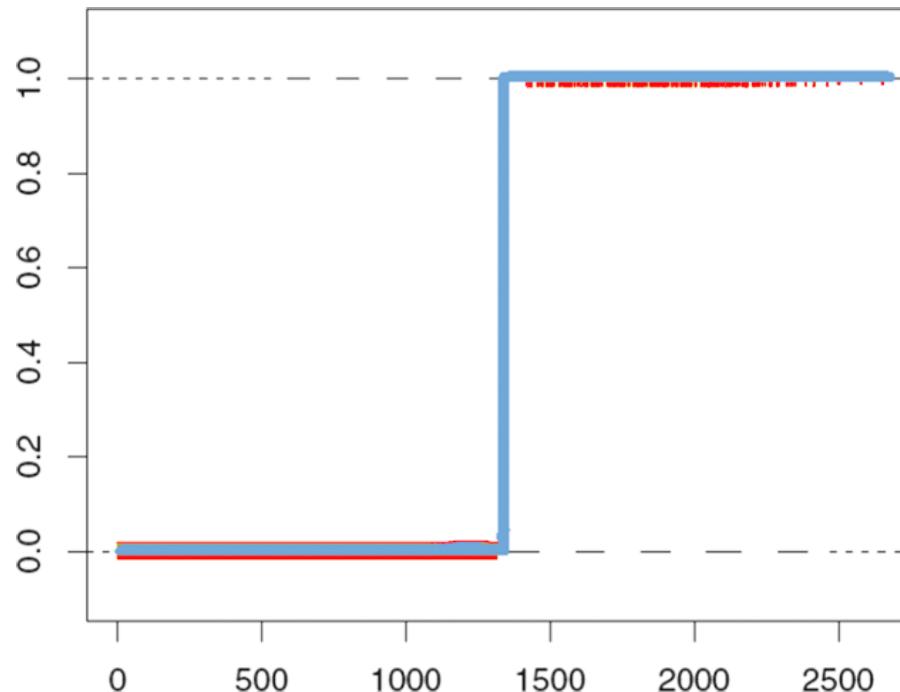
$$\ln L(X) = \sum_{i=1}^{\ell} \left( y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right)$$

- » Хорошо максимизируется численно  
(задача выпуклая)

# ОЦЕНКА ПАРАМЕТРОВ

$$\ln L(X) = \sum_{i=1}^{\ell} \left( y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right)$$

» Если  $y = 0$  и  $y = 1$  линейно разделимы в пространстве признаков, то оптимизация приводит к  $\| w \| \rightarrow \infty$



# ОЦЕНКА ПАРАМЕТРОВ

---

$$\ln L(X) = \sum_{i=1}^{\ell} \left( y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i)) \right)$$

- » Если  $y = 0$  и  $y = 1$  линейно разделимы в пространстве признаков, то оптимизация приводит к  $\| w \| \rightarrow \infty$
- » Решение — регуляризация

# ПРЕДСКАЗАНИЕ ОТКЛИКА

---

- » Имея оценку  $\pi(x)$ , можно предсказывать  
 $y = 1$  при  $\pi(x) > p_0$
- » Варианты:
  - ▶  $p_0 = 0.5$
  - ▶ подбор  $p_0$  для достижения нужного баланса между точностью и полнотой классификатора

# РЕЗЮМЕ

---

- › Логистическая регрессия позволяет предсказывать вероятность  $y = 1$  по  $x$
- › Используется линейная модель логита
$$\ln \frac{P(y = 1|x)}{P(y = 0|x)}$$
- › Оценка параметров делается методом максимального правдоподобия

# ДАЛЕЕ В ПРОГРАММЕ

---

- › Технические трюки в линейной регрессии