Системы обработки медиаданных

Линейная регрессия

д.т.н., доцент Вашкевич М. И.

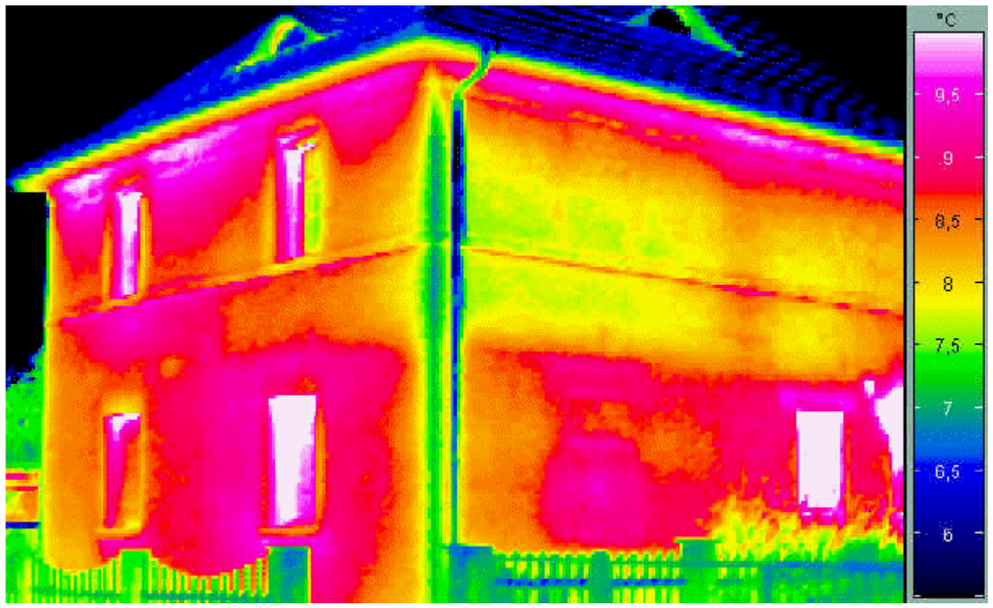
vashkevich@bsuir.by

Белорусский государственный университет информатики и радиоэлектроники

Кафедра электронных вычислительных средств

Пример

Как потребление газа зависит от внешней температуры? (Whiteside, 1960-е)



Еженедельно измерялась:

– средняя температура воздуха

– коэффициент теплоизоляции

– общий объем использованного газа

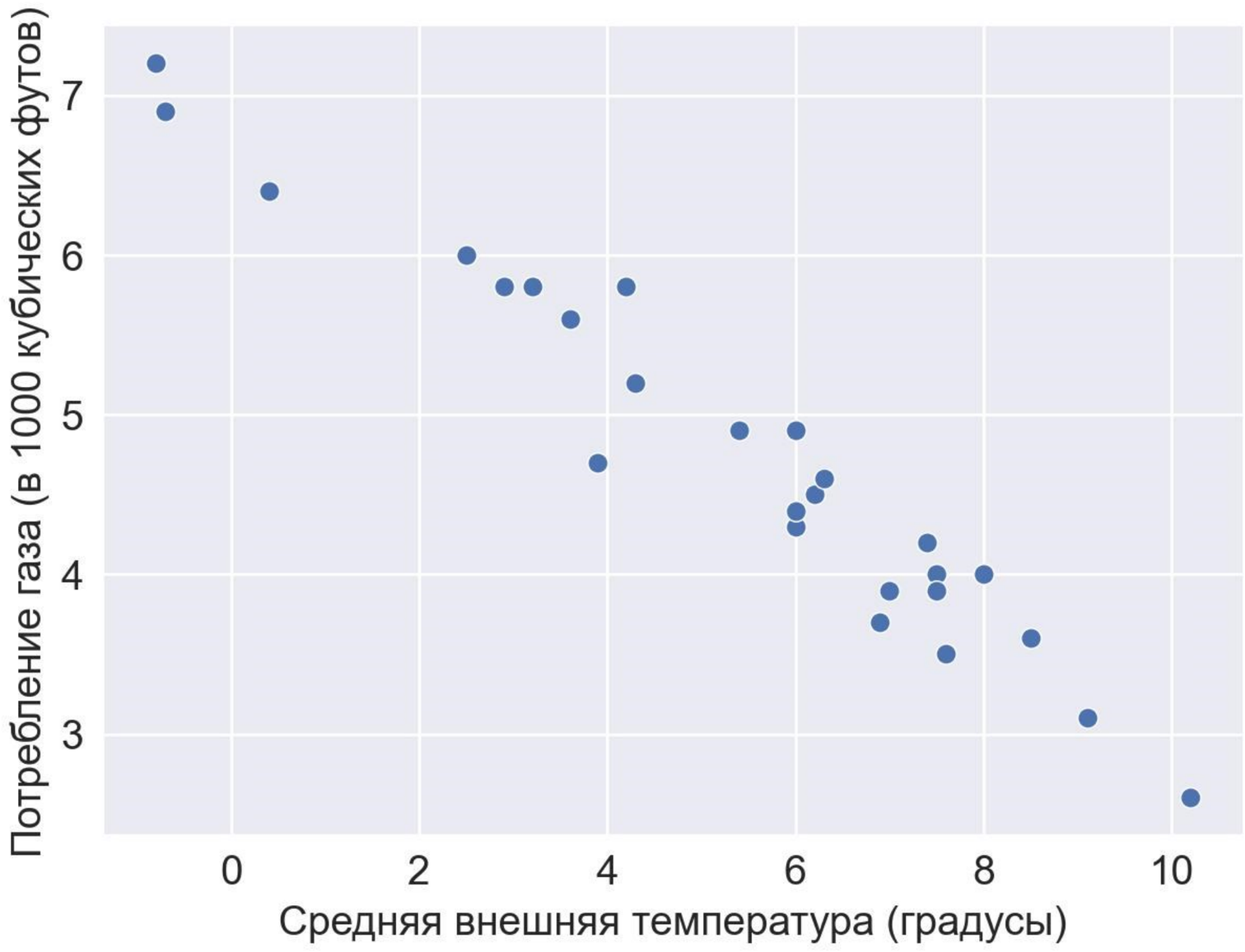
Набор данных 56 строк и 3 столбца:

Вопросы:

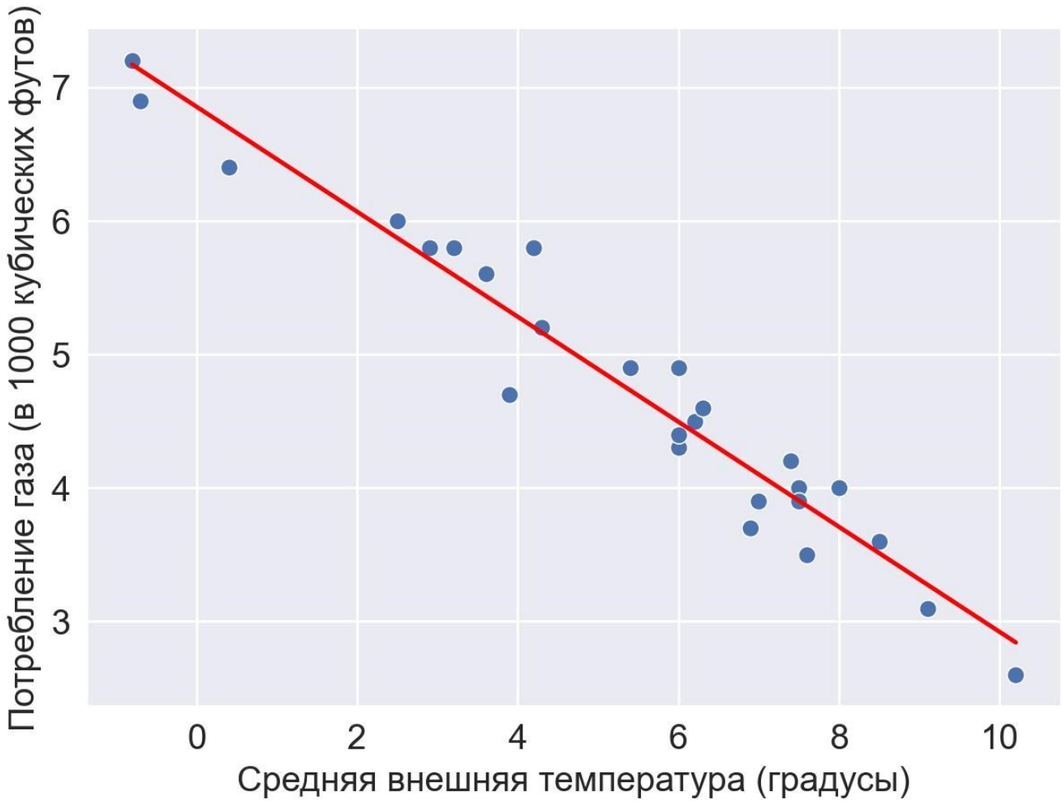
– Как потребление газа зависит от внешней температуры воздуха?

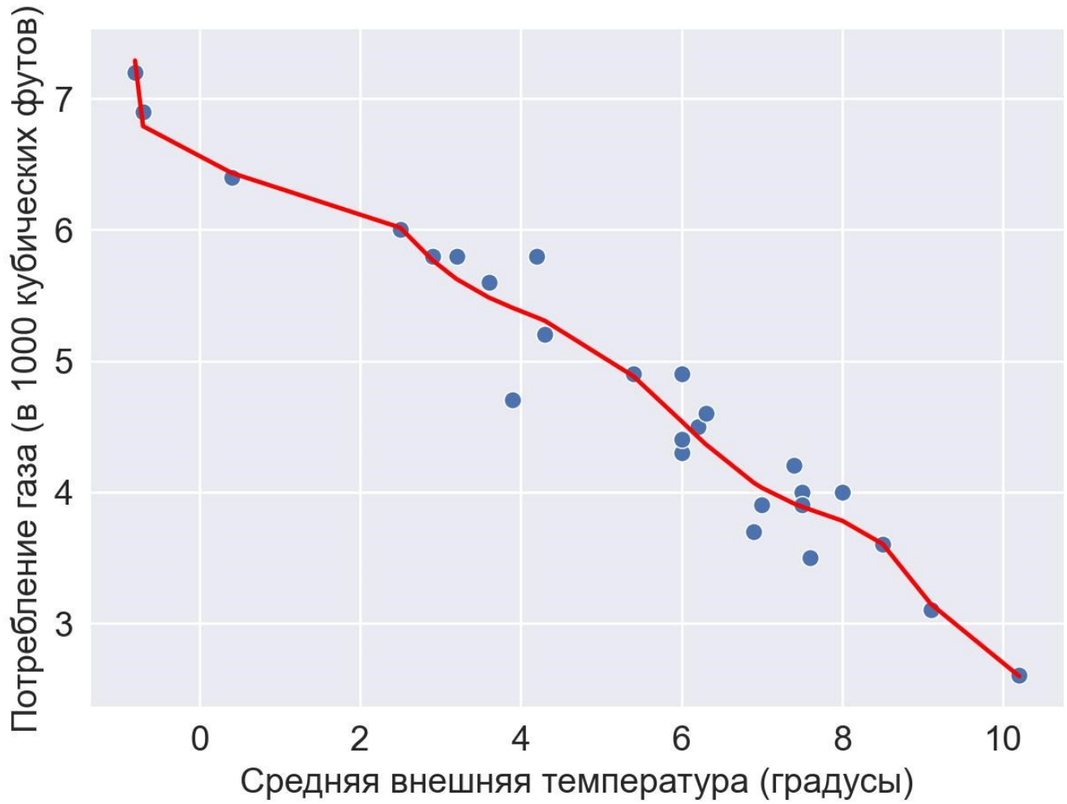
– Сколько газа необходимо при заданной температуре?

Пример


Потребление газа (в 1000 кубических футов)
7
6
5
4
3
0
2
4
6
8
10
Средняя внешняя температура (градусы)

Пример





Линейная модель

Более гибкая модель

Обозначения

• Обычно в нашем распоряжении есть выборка, содержащая n наблюдений.

• Каждое наблюдение состоит из p переменных (или параметров), на основании которых мы делаем предсказания.

• При помощи x_{ij} – обозначают i-е значение, j-й переменной, где $i = 1,2, \dots, n$, а $j = 1,2, \dots, p$.

• При помощи $\mathbf{X}$ будем обозначать матрицу $n \times p$, чей (i, j)-й элемент это x_{ij}:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Обозначения

• Строки матрицы $\mathbf{X}$ будем записывать, как $x_1, x_2, \ldots, x_n$. Здесь x_i – вектор длиной p переменных i-го наблюдения:

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}.$$

• Столбцы матрицы $\mathbf{X}$ будем записывать, как $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. Здесь $\mathbf{x}_j$ – вектор длиной n j-го признака всех наблюдений:

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

• Матрицу $\mathbf{X}$ можно записать, как $\mathbf{X} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_p)$.

Формальная постановка задачи регрессии

Пусть

$X_1,\ X_2,\ X_p$ – случайные величины называемые ***предикторами*** (или ***входами***, ***независимыми переменными***).

Пусть $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$ – их область определения.

Будем использовать короткую форму записи

$$X \coloneqq (X_1, X_2, \ldots, X_p)$$

для вектора предикторов и

$$\mathcal{X} \coloneqq \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_p$$

для области определения

Y – случайная величина, называемая ***целевой переменной*** (или ***выход***, ***отклик***)

Пусть $\mathcal{Y}$ – область определения.

$\mathcal{D} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ – множество примеров из неизвестного совместного распределения $p(X, Y)$ входов и выходов, которое называется ***данными***.

$\mathcal{D}$ обычно записывается списком

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

Формальная постановка задачи регрессии

- Задача регрессии (и классификации) это **предсказать** Y **на основе** X.

- Другими словами, необходимо оценить функцию

$$f(x) \coloneqq E\{Y|X = x\} = \int y \cdot p(y|x)dy$$

на основе данных. $f(x)$ называется ***функцией регрессии***.

- Если Y принимает численные значения, то задача называется **регрессией**.

- Если Y принимает номинальные значения, то задача называется **классификацией**.

Формальная постановка задачи регрессии

• Мы предполагаем, что между Y и X существует связь, которую можно записать в общей форме:

$$Y = f(X) + \epsilon,$$

f – фиксированная, но неизвестная функция от $X_1, X_2, \ldots, X_p$, ϵ – ошибка, которая не зависит от X.

• Функция f выражает ***систематическую*** информацию о Y, содержащуюся в X.

Зачем оценивать f?

1) Предсказание Y:

$$\hat{Y} = \hat{f}(X)$$

$\hat{f}$ – оценка f, $\hat{Y}$ – предсказанное значение Y.

2) Статистический вывод.

Мы можем найти ответы на вопросы: какие предикторы связаны с откликом? Какова связь между откликом и каждым предиктором? и т.д.

Предсказание

• Точность $\hat{Y}$ в качестве предсказанного значения Y зависит от:
1) ***устранимой ошибки*** и 2) ***неустранимой ошибки***.

• Устранимая ошибка связана с неидеальной оценкой f.

• Неустранимая ошибка связана с наличием ϵ (шума), который возникает из-за неучтенных факторов.

Пусть $\hat{Y} = \hat{f}(X)$, причем $\hat{f}$ и X являются фиксированными, тогда:

$$E\left\{\left(Y - \hat{Y}\right)^2\right\} = E\left\{\left(f(X) + \epsilon - \hat{f}(X)\right)^2\right\}$$

$$= \underbrace{\left(f(X) - \hat{f}(X)\right)^2}_{\text{устранимая}} + \underbrace{\text{Var}\{\epsilon\}}_{\text{неустранимая}}.$$

Модель простой линейной регрессии

– Самый простой предиктор X это одномерный вектор, т.е. ($X = X_1$)

– $f(x)$ – предполагается линейной:

$$f(x) = w_0 + w_1 x$$

– модель линейной регрессии предполагает, что

$$y_i = w_0 + w_1 x_i + \varepsilon_i, \qquad E\{\varepsilon\} = 0, \qquad Var\{\varepsilon\} = \sigma^2.$$

– 3 параметра линейной регрессии

w_0 – **пересечение** (свободный член)
w_1 – **угловой коэффициент** (коэффициент регрессии)
σ^2 – **дисперсия остатков**

Модель простой линейной регрессии

Оценки параметров

$$\widehat{w}_0, \widehat{w}_1, \hat{\sigma}^2$$

Подогнанная линейная функция

$$\hat{f}(x) = \widehat{w}_0 + \widehat{w}_1 x$$

Предсказанные / подогнанные значения

$$\hat{y}_i = f(x_i)$$

Остатки / отклонения

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (\widehat{w}_0 + \widehat{w}_1 x_i)$$

Сумма квадратов остатков

$$RSS = \sum_{i=1}^{n} {\varepsilon_i}^2$$

Как оценить параметры?

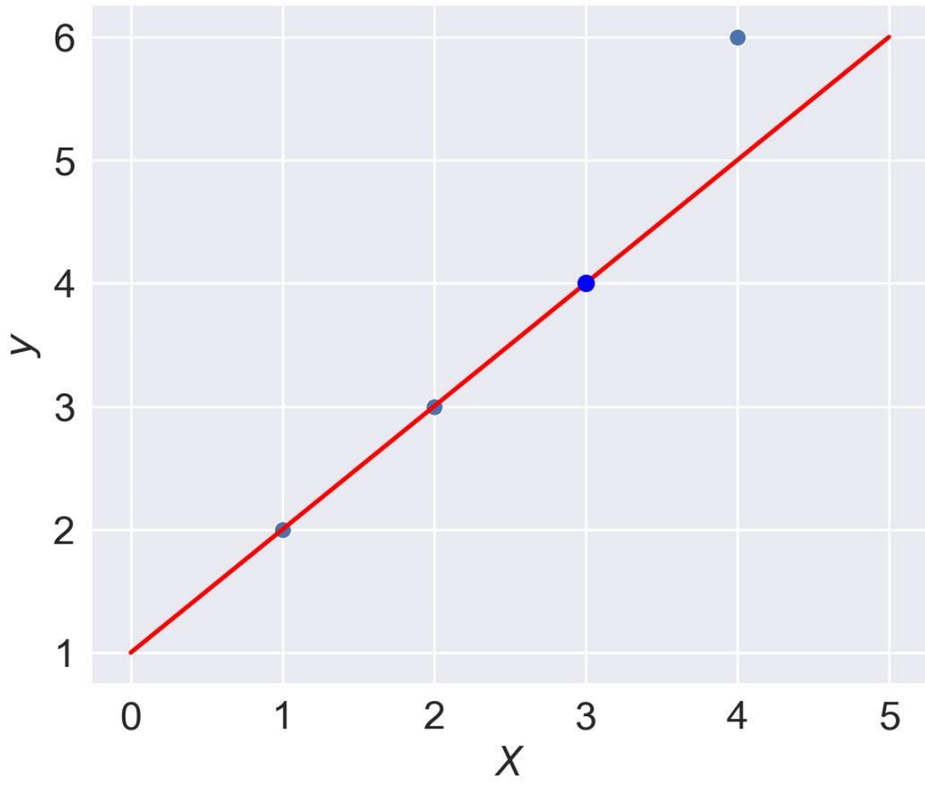
Пример

Используя данные $\mathcal{D} := \{(1,2), (2,3), (4,6)\}$, предсказать значения для $x = 3$.

Построим линию через первую и вторую точку

$$\widehat{w}_1 = \frac{y_2 - y_1}{x_2 - x_1} = 1$$

$$\widehat{w}_0 = y_1 - \widehat{w}_1 x_1 = 1$$



RSS:

i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$
1	2	2	0
2	3	3	0
3	6	5	1
Σ			1

$$\hat{r}(3) = 4$$

Как оценить параметры?

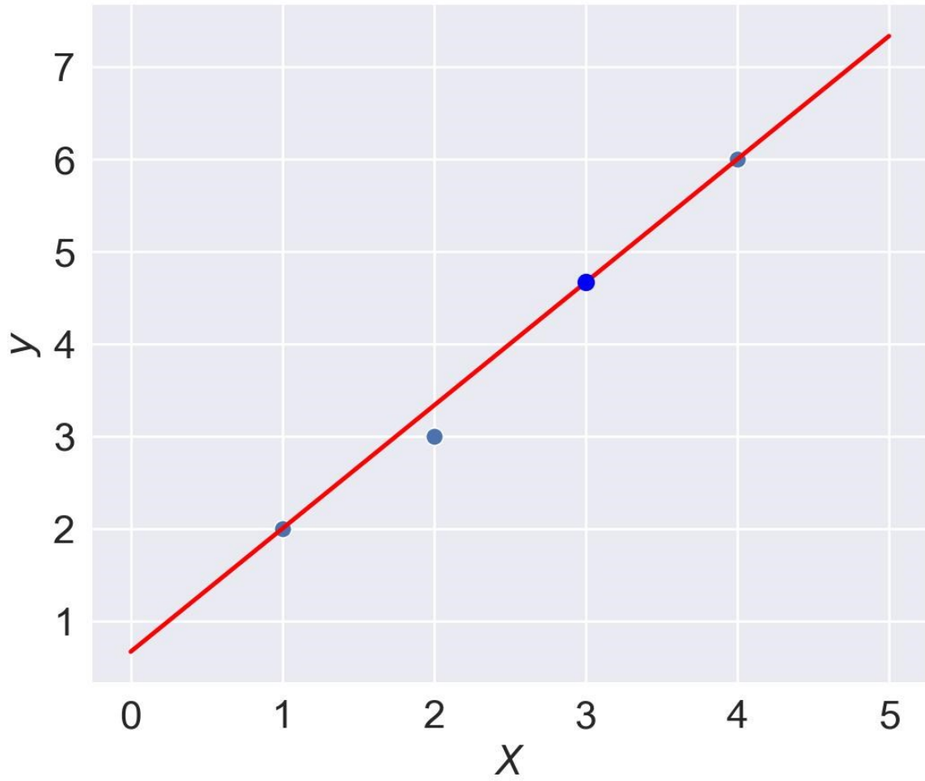
Пример

Используя данные $\mathcal{D} := \{(1,2), (2,3), (4,6)\}$, предсказать значения для $x = 3$.

Построим линию через первую и последнюю точку

$$\widehat{w}_1 = \frac{y_3 - y_1}{x_3 - x_1} = \frac{4}{3} = 1.333$$

$$\widehat{w}_0 = y_1 - \widehat{w}_1 x_1 = 2/3$$



RSS:

i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$
1	2	2	0
2	3	3.333	0.333
3	6	6	0
Σ			0.333

$$\hat{r}(3) = 4.667$$

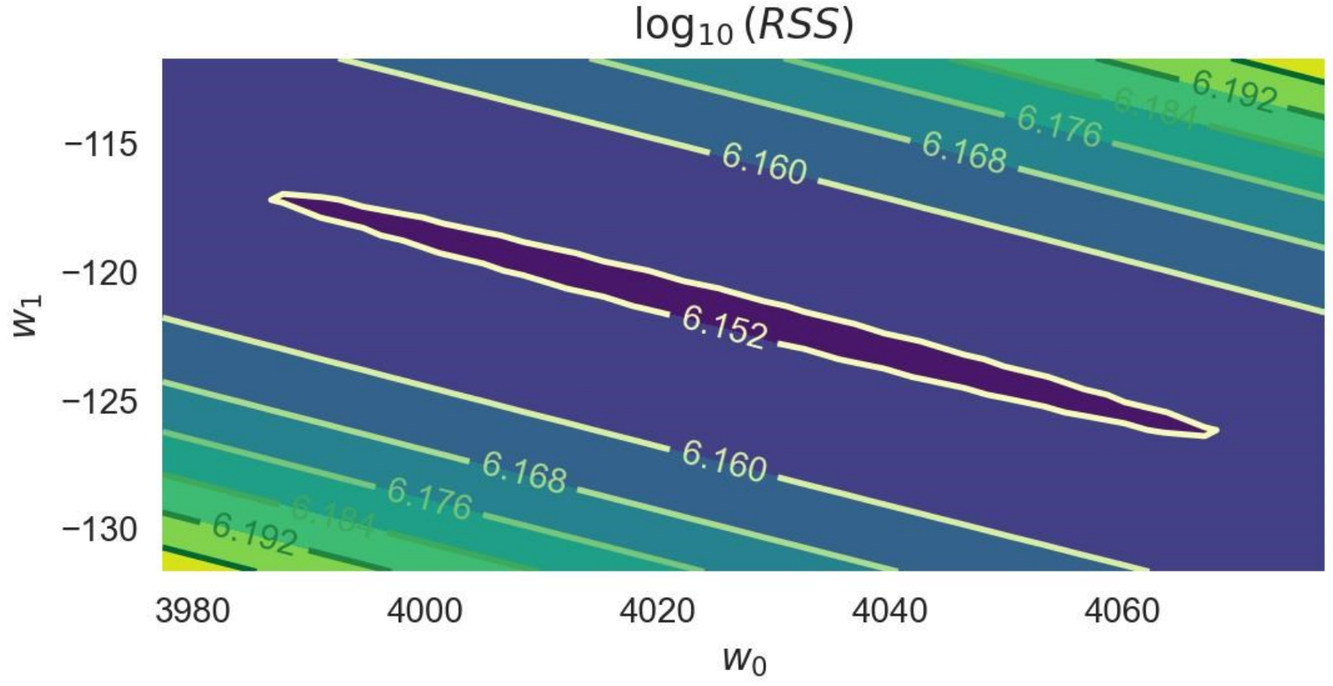
Метод наименьших квадратов

Существует много различных способов оценки по данным параметров $\hat{w}_0, \hat{w}_1$ и $\hat{\sigma}^2$, которые отличаются свойствами, которыми в результате будет обладать решение.

Рассмотрим функцию

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 \quad = \sum_{i=1}^{n} (y_i - (\hat{w}_0 + \hat{w}_1 x_i))^2.$$

для некоторого набора данных

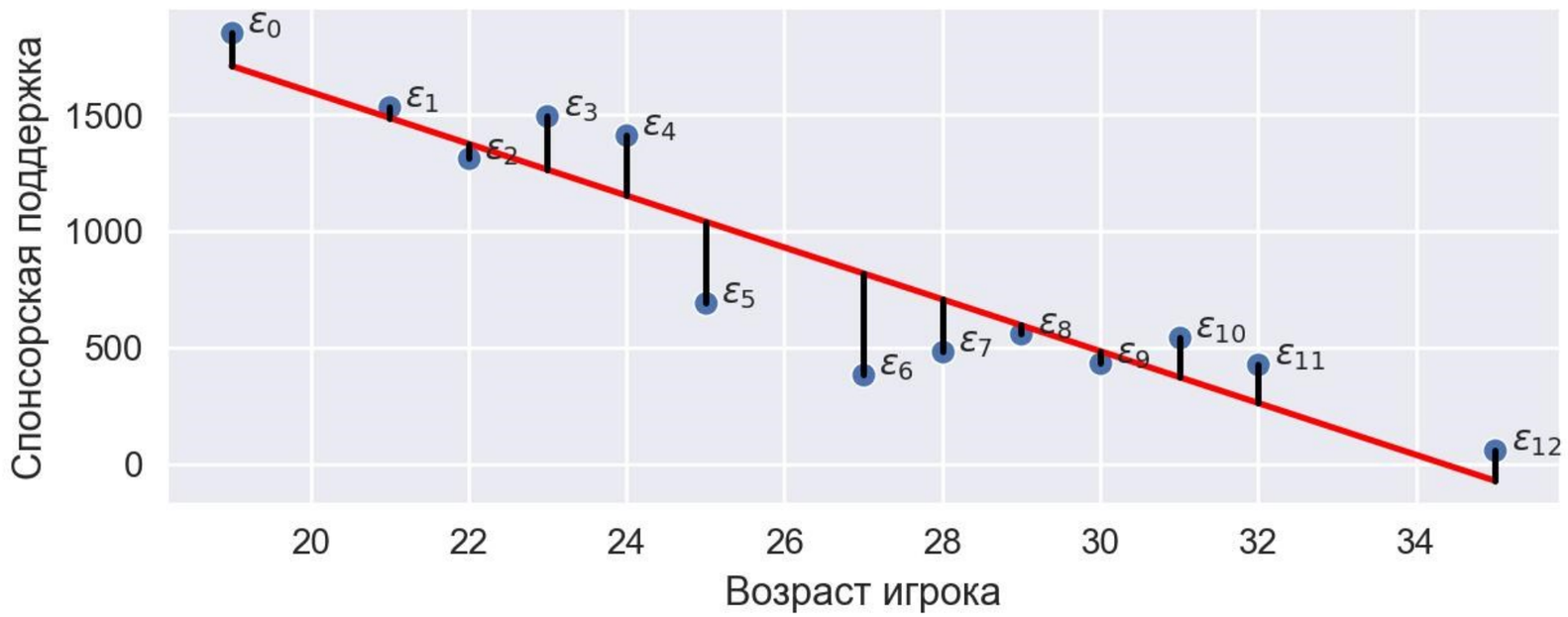


Метод наименьших квадратов

Существует много различных способов оценки по данным параметров $\hat{w}_0$, $\hat{w}_1$ и $\hat{\sigma}^2$, которые отличаются свойствами, которыми в результате будет обладать решение.

Оценка методом наименьших квадратов (*LS – least square*) этих параметров осуществляется путем минимизации

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 \quad = \sum_{i=1}^{n} (y_i - (\hat{w}_0 + \hat{w}_1 x_i))^2.$$



Метод наименьших квадратов

Метод наименьших квадратов осуществляется путем минимизации

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 \quad = \sum_{i=1}^{n} (y_i - (\hat{w}_0 + \hat{w}_1 x_i))^2.$$

Найдем производную по $\hat{w}_0$:

$$\frac{\partial RSS}{\partial \hat{w}_0} = \sum_{i=1}^{n} 2(y_i - \hat{w}_0 - \hat{w}_1 x_i)(-1) \triangleq 0$$

$$\Rightarrow \hat{w}_0 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{w}_1 x_i)$$

$$\hat{w}_0 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{w}_1 x_i) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} y_i}_{\bar{y}} - \hat{w}_1 \underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i}_{\bar{x}} = \bar{y} - \hat{w}_1 \bar{x}.$$

Метод наименьших квадратов

Перепишем RSS учитывая выражение для $\widehat{w}_0$

$$RSS = \sum_{i=1}^{n}(y_i - (\widehat{w}_0 + \widehat{w}_1 x_i))^2 = \sum_{i=1}^{n}(y_i - (\bar{y} - \widehat{w}_1 \bar{x}) - \widehat{w}_1 x_i)^2$$

$$= \sum_{i=1}^{N}(y_i - \bar{y} - \widehat{w}_1 (x_i - \bar{x}))^2.$$

Найдем производную по $\widehat{w}_1$:

$$\frac{\partial RSS}{\partial \widehat{w}_1} = \sum_{i=1}^{n} 2\big(y_i - \bar{y} - \widehat{w}_1 (x_i - \bar{x})\big)(-1)(x_i - \bar{x}) \triangleq 0.$$

$$\Rightarrow \widehat{w}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Метод наименьших квадратов

Оценка методом наименьших квадратов параметров $\hat{w}_0, \hat{w}_1$ и $\hat{\sigma}^2$ осуществляется путем минимизации

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 \ = \sum_{i=1}^{n} (y_i - (\hat{w}_0 + \hat{w}_1 x_i))^2.$$

Решение можно записать в замкнутом виде:

$$\hat{w}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2$$

Метод наименьших квадратов (МНК) / Пример

Пример

Используя данные $\mathcal{D} := \{(1,2), (2,3), (4,6)\}$, предсказать значения для $x = 3$.

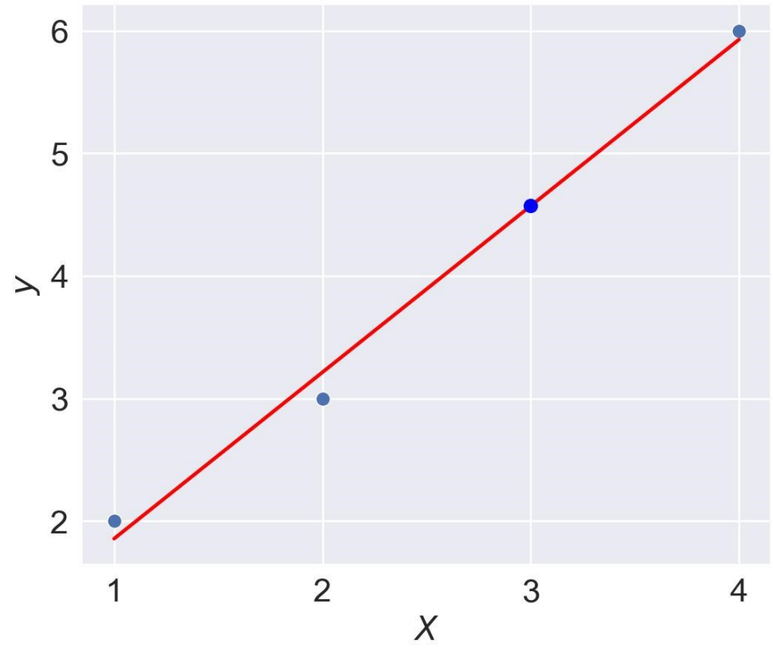
Построим линию регрессии методом МНК.

$\bar{x} = 7/3, \bar{y} = 11/3.$

i	$x_i - \bar{x}$	$y_i - \hat{y}_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	$-4/3$	$-5/3$	$16/9$	$20/9$
2	$-1/3$	$-2/3$	$1/9$	$2/9$
3	$5/3$	$7/3$	$25/9$	$35/9$
Σ			$42/9$	$57/9$

$$\hat{w}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{57}{42} \approx 1{,}36$$

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x} = \frac{11}{3} - \frac{57}{42} \cdot \frac{7}{3} = 0{,}5$$



Метод наименьших квадратов (МНК) / Пример

Пример

Используя данные $\mathcal{D} := \{(1,2), (2,3), (4,6)\}$, предсказать значения для $x = 3$.

Построим линию регрессии методом наименьших квадратов.

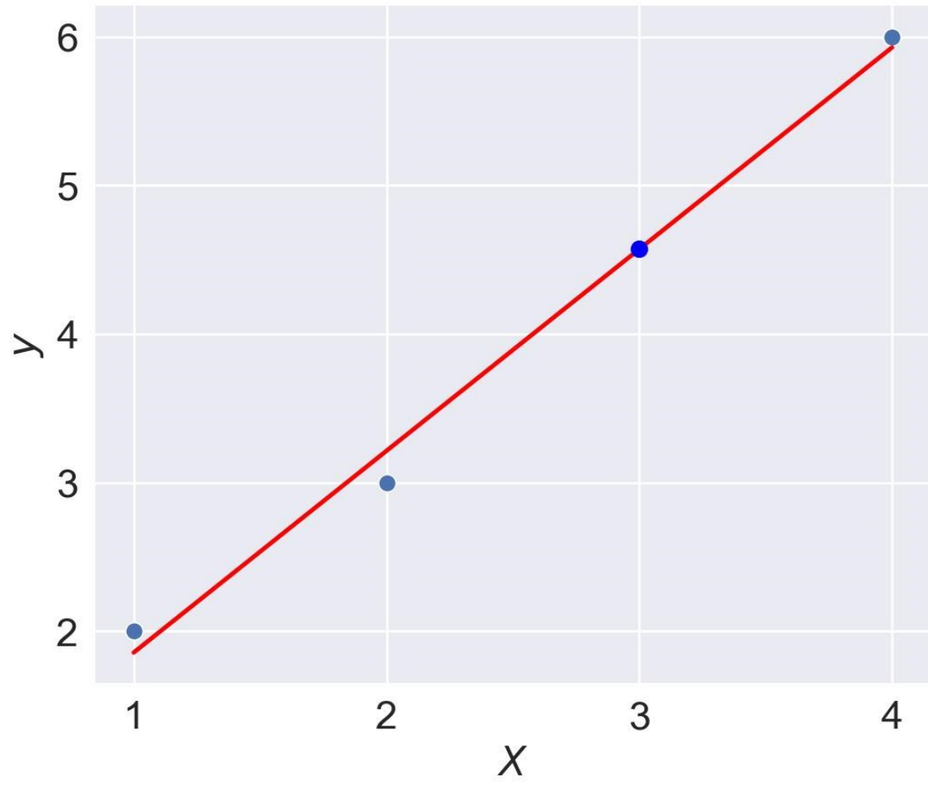
$$\widehat{w}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{57}{42} \approx 1{,}36$$

$$\widehat{w}_0 = \bar{y} - \widehat{w}_1 \bar{x} = \frac{11}{3} - \frac{57}{42} \cdot \frac{7}{3} = 0{,}5$$

RSS:

i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	2	1,86	0,020
2	3	3,21	0,050
3	6	5,93	0,005
Σ			0,071

$$\hat{r}(3) = 4{,}571$$



Особенности уравнения линейной регрессии

Оценка по методу наименьших квадратов

Свободный член

$$\widehat{w}_0 = \bar{y} - \widehat{w}_1 \bar{x}$$

Коэффициент регрессии

$$\widehat{w}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Коэффициент регрессии содержит умноженную на n ковариацию между X и Y и умноженную на n дисперсию X в знаменателе

$$\widehat{w}_1 = \frac{n \cdot \mathrm{Cov}(X, Y)}{n \cdot \mathrm{Var}\{X\}} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}\{X\}}.$$

Особенности уравнения линейной регрессии

Уравнение линейной регрессии (МНК):

$$\hat{y} = \underbrace{\bar{y} - \hat{w}_1\bar{x}}_{\hat{w}_0} + \hat{w}_1 x = \bar{y} + \hat{w}_1(x - \bar{x})$$

Решение проходит через точку $(\bar{x}, \bar{y})$ ⇒ можно выполнить центрирование:

$$\check{x}_i = x_i - \bar{x}$$

Оценка по методу наименьших квадратов (центрированные данные)

Свободный член

$$\hat{w}_0 = \bar{y}$$

Коэффициент регрессии

$$\hat{w}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\check{x}_i)}{\sum_{i=1}^{n}(\check{x}_i)^2}$$

Нормирование и нормализация данных

Нормирование означает

$$x_i' = \frac{x_i}{\sigma_x} \quad \Longrightarrow \quad \bar{x}' = \frac{\bar{x}}{\sigma_x}$$

Оценка по методу наименьших квадратов (нормированные данные)	Свободный член	Коэффициент регрессии
	$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}'$	$\hat{w}_1 = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}')$

Нормализация = центрирование относительно 0 + нормирование

$$x_i' = \frac{x_i - \bar{x}}{\sigma_x}$$

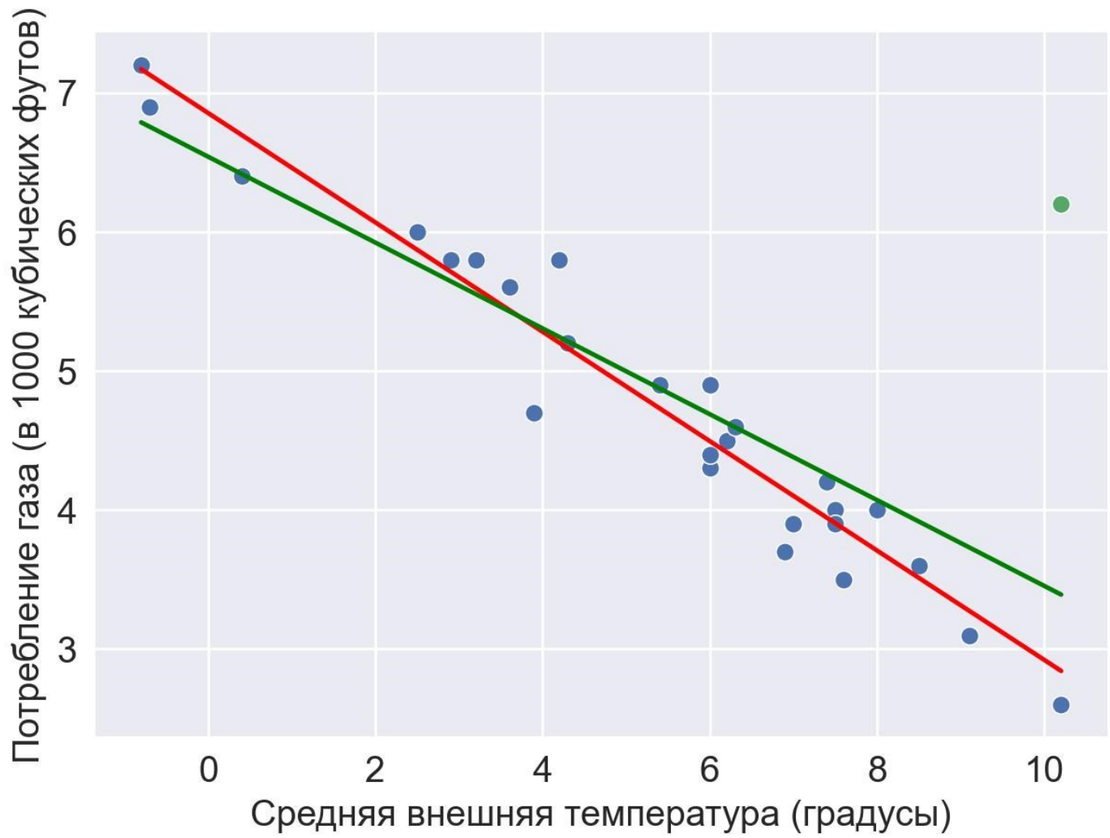
Оценка по методу наименьших квадратов (нормализованные данные)	Свободный член	Коэффициент регрессии
	$\hat{w}_0 = \bar{y}$	$\hat{w}_1 = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}')$

Линейная регрессия (МНК): анализ отклонений

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - (\hat{w}_0 + \hat{w}_1 x_i) = 0$$

Это следует из того, что $\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$. [Доказать].

Эта особенность делает линейную регрессию чувствительной к **выбросам**.



Векторные обозначения

Обозначим через $\boldsymbol{x}_i = [1 \;\; x_i]^T$, а через $\mathbf{w} = [w_0 \;\; w_1]^T$, тогда уравнение линейной регрессии перепишется, как

$$y_i = \mathbf{x}_i^T \mathbf{w} + \epsilon_i$$

Такая запись особенно удобна, когда число предикторов больше одного.

Множественная регрессия

Несколько предикторов $X_1, X_2, \dots, X_p$

$$Y = w_0 + w_1 X_1 + w_2 X_2 + \cdots + w_p X_p + \varepsilon =$$
$$= \mathbf{X}^T \mathbf{w} + \varepsilon$$

с $p + 1$ параметрами $\mathbf{w} = [w_0, w_1, \dots, w_p]$.

Матричная запись (линейная форма)

Несколько предикторов $X_1, X_2, \ldots, X_p$

$$Y = w_0 + \sum_{i=1}^{p} w_i X_i + \varepsilon$$
$$= \mathbf{X}^T \mathbf{w} + \varepsilon$$

где

$$\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{pmatrix}, \quad \mathbf{X} := \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{pmatrix}$$

свободный член (w_0) рассматривается, как и все остальные параметры, но для искусственно заданной константной переменной $X_0 = 1$.

Уравнение для всего набора данных

Для всего набора данных $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ \vdots \\ w_d \end{pmatrix}}_{\mathbf{w}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}.$$

т.е.

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \varepsilon$$

Оценка по МНК

Оценка $\widehat{\boldsymbol{w}}$ по методу наименьших квадратов (МНК) минимизирует:

$$RSS(\widehat{w}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - \mathbf{X}\widehat{w}\|^2$$

МНК оценка вектора $\widehat{\boldsymbol{w}}$ вычисляется по формуле:

$$\widehat{\boldsymbol{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Доказательство: ?

Оценка по МНК

Оценка $\widehat{\boldsymbol{w}}$ по методу наименьших квадратов (МНК) минимизирует:

$$RSS(\hat{w}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{w}\|^2$$

МНК оценка вектора $\hat{w}$ вычисляется по формуле:

$$\widehat{\boldsymbol{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Доказательство:

$$RSS(\widehat{\boldsymbol{w}}) = (\mathbf{y} - \mathbf{X}\hat{w})^T(\mathbf{y} - \mathbf{X}\hat{w})$$

Это квадратичная функция с $p + 1$ параметрами.

$$\frac{\partial RSS}{\partial \hat{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{w}}), \quad \frac{\partial^2 RSS}{\partial \hat{w}\partial \hat{w}^T} = 2\mathbf{X}^T\mathbf{X}$$

$$-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{w}) = 0 \quad \Rightarrow \quad \widehat{\boldsymbol{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Множественная регрессия

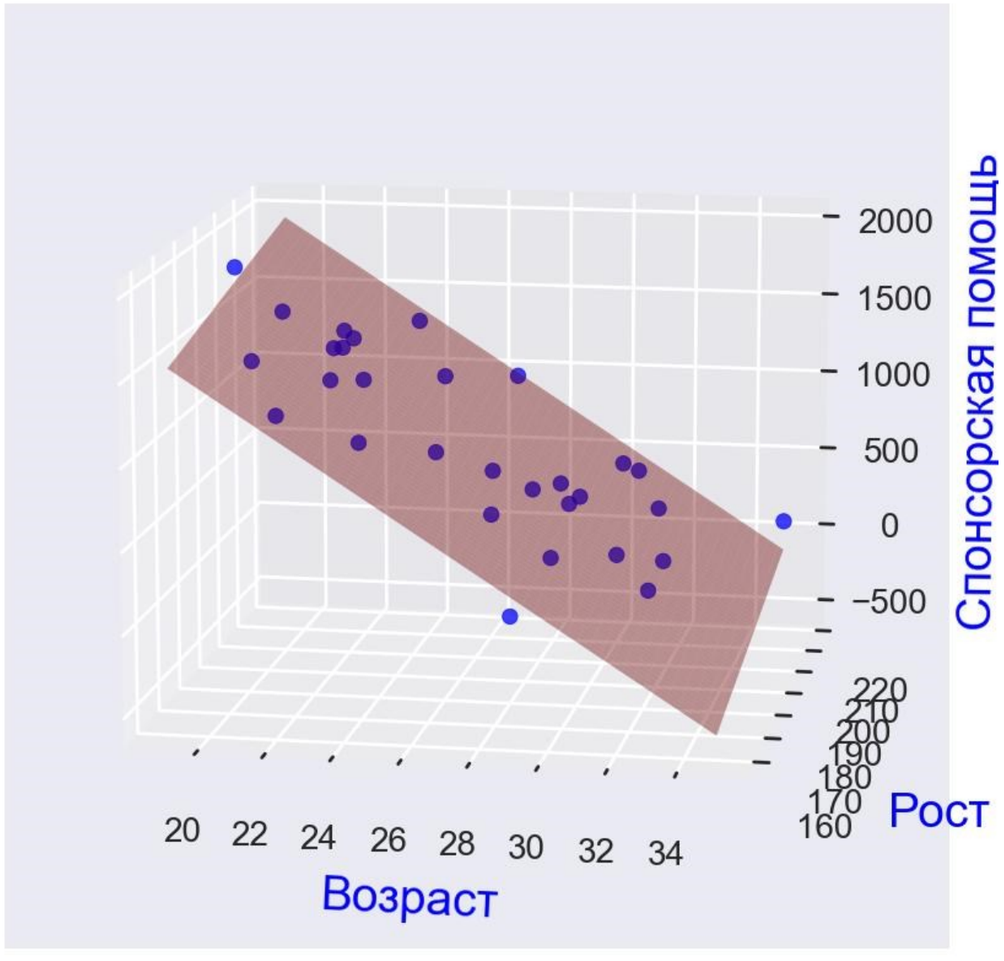
Предположим, у нас есть данные о игроках баскетбольной команды, которые включают в себя такие параметры, как рост, вес, возраст и средненедельный объем спонсорской поддержки.

#	Возраст (X_1), лет	Рост (X_2), см	Вес (X_3), фунты	Спонсорская помощь (Y), $
1	29	192	218	561
2	35	218	251	60
3	22	197	221	1312
4	22	192	219	1359
5	29	198	223	362
6	21	166	188	1536
…	…	…	…	…

Множественная регрессия: пример

Пример с использованием данных о игроках баскетбольной команды.

Предикторы: возраст, рост; **целевая переменная**: средненедельный объем спонсорской поддержки.



Градиентный спуск

Обучение = минимизация функции потерь.

- Минимизация осуществляется методом градиентного спуска

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \frac{\partial RSS}{\partial \mathbf{w}},$$

где $\mathbf{w}_t$ – значение весов на шаге t, $\frac{\partial RSS}{\partial \mathbf{w}}$ – градиент функции потерь по $\mathbf{w}$, η – параметр скорости обучения.

- Выбор параметра η – значительно влияет на процесс обучения.

Линейная регрессия, как генеративная модель

Распределение Гаусса

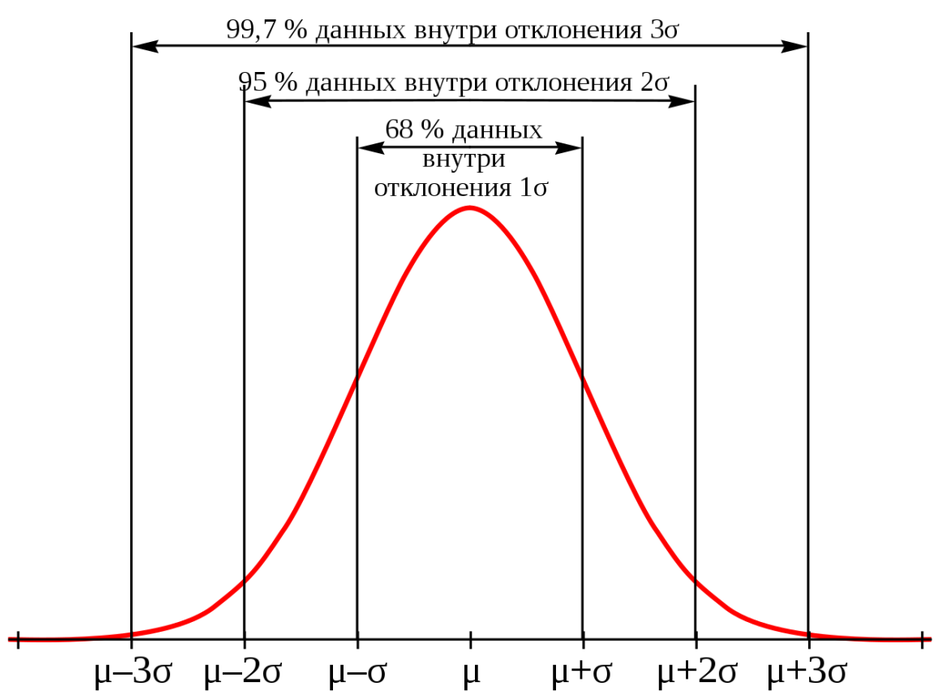
• Случайная величина x с нормальным распределением принимает значения в интервале $(-\infty, \infty)$ и имеет функцию ***плотности вероятности***

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

где μ – математическое ожидание, σ^2 – дисперсия.

Правило трех сигм

Вероятность того, что случайная погреш-ность отдельного анализа не превысит по абсолютному значению σ (или 2σ, 3σ), составляет 0,68 (или 0,95 , 0,99).



Стандартное нормальное распределение

• Гауссово распределение $p(x)$ обозначают, как $\mathcal{N}(x \mid \mu, \sigma^2)$ или $X \sim \mathcal{N}(\mu, \sigma^2)$

• Все нормальные распределения могут быть сведены к одному нормальному распределению с параметрами $\mu = 0, \sigma^2 = 1$. Для этого нужно выполнить преобразование

$$u = \frac{x - \mu}{\sigma}.$$

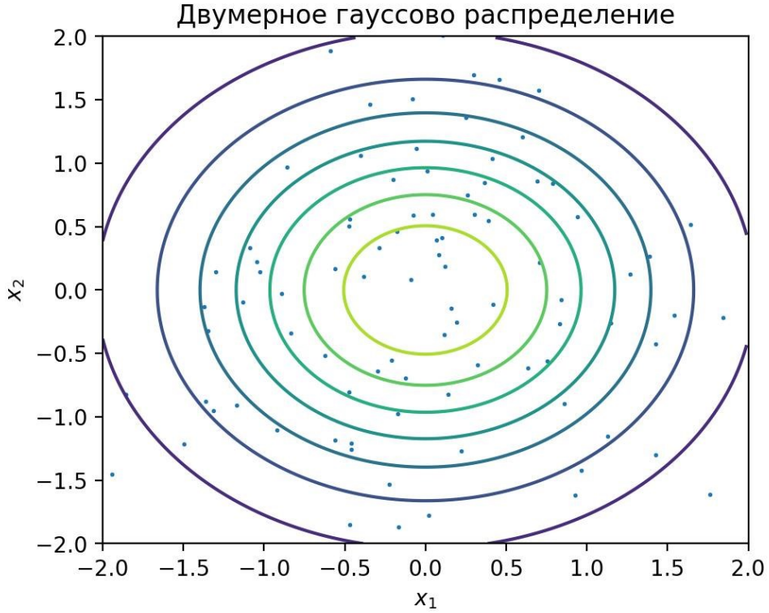
• Если с.в. $X \sim \mathcal{N}(\mu, \sigma^2)$, то $U = \frac{X-\mu}{\sigma}$ имеет распределение $U \sim \mathcal{N}(0,1)$.

Многомерное гауссово распределение

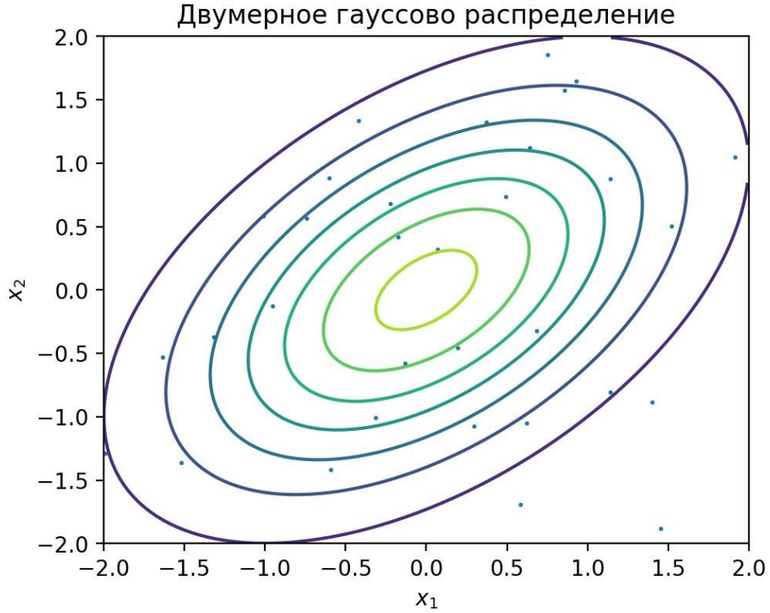
Многомерное гауссово распределение полностью определяется вектором средних $\boldsymbol{\mu}$ и ковариационной матрицей $\boldsymbol{\Sigma}$

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} \cdot |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ – короткая запись, D – размерность вектора $\boldsymbol{x}$.



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Генеративная модель

Модель линейной регрессии предполагает, что

$$y_i = \mathbf{x}_i^T \mathbf{w} + \varepsilon_i, \qquad E\{\varepsilon\} = 0, \qquad \mathrm{Var}\{\varepsilon\} = \sigma^2.$$

Мы накладываем ограничения на отклонения ε, но не задаем закона их распределения.

Мы можем сделать предположение относительно распределения ε:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

следовательно

$$Y \sim \mathcal{N}(\mathbf{X}^T \mathbf{w}, \sigma^2).$$

Мы можем сгенерировать выборку по этой модели.

Метод максимального правдоподобия

Пусть $p(X, Y|\mathbf{w})$ совместная плотность распределения X и Y, а $\mathbf{w}$ – неизвестный параметр(ы) распределения.

Функция ***правдоподобия***

$$L_{\mathcal{D}}(\mathbf{w}) \coloneqq \prod_{i=1}^{n} p(x_i, y_i|\mathbf{w}).$$

Правдоподобие показывает вероятность появления набора данных $\mathcal{D}$.

Метод максимального правдоподобия

Пусть $p(X, Y|\mathbf{w})$ совместная плотность распределения X и Y, а $\mathbf{w}$ – неизвестный параметр(ы) распределения.

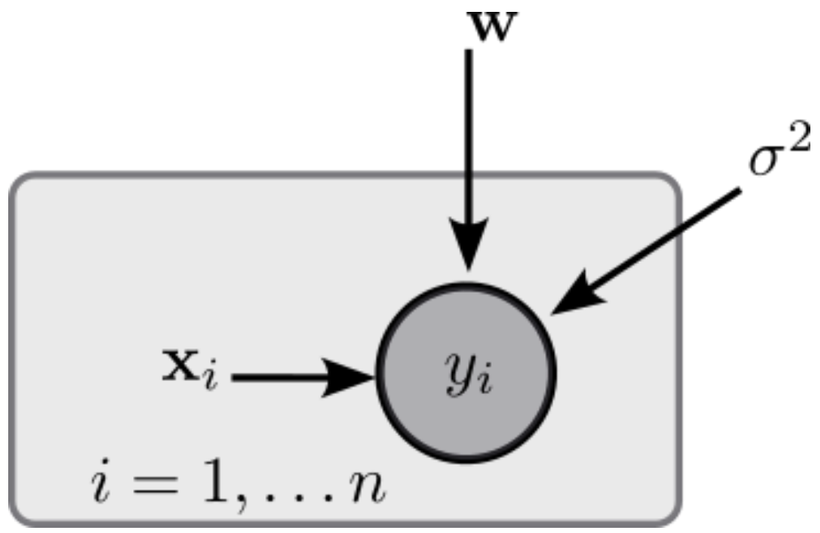
Функция ***правдоподобия***

$$L_{\mathcal{D}}(\mathbf{w}) := \prod_{i=1}^{n} p(x_i, y_i|\mathbf{w}).$$

Правдоподобие показывает вероятность появления набора данных $\mathcal{D}$.

Сущность метода ***максимального правдоподобия*** состоит в том, чтобы найти такое значение параметра $\mathbf{w}$, которое будет максимизировать $L_{\mathcal{D}}(\mathbf{w})$. При максимизации подразумевается, что данные $\mathcal{D}$ фиксированы, а изменяется только параметр $\mathbf{w}$:

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} L_{\mathcal{D}}(\mathbf{w})$$



Применение метода максимального правдоподобия

• При оценивании неизвестных параметров распределения на практике часто используют метод максимального правдоподобия (*MLE – maximum likelihood estimation*).

Применение метода максимального правдоподобия

• При оценивании неизвестных параметров на практике часто используют метод максимального правдоподобия (*MLE – maximum likelihood estimation*).

• Пусть есть случайные величины $x_1, x_2, \ldots, x_n$ каждая из которых распределены по закону $p(x|\theta)$. Предполагается, что функция $p(x|\theta)$ известная с точностью до параметра θ.

Применение метода максимального правдоподобия

• При оценивании неизвестных параметров на практике часто используют метод максимального правдоподобия (*MLE – maximum likelihood estimation*).

• Пусть есть случайные величины $x_1, x_2, \ldots, x_n$ каждая из которых распределены по закону $p(x|\theta)$. Предполагается, что функция $p(x|\theta)$ известная с точностью до параметра θ.

• Функция

$$L(\theta) \coloneqq \prod_{i=1}^{n} p(x_i|\theta).$$

представляет собой случайную величину и называется ***функцией правдоподобия***.

Применение метода максимального правдоподобия

• При оценивании неизвестных параметров на практике часто используют метод максимального правдоподобия (*MLE – maximum likelihood estimation*).

• Пусть есть случайные величины $x_1, x_2, \dots, x_n$ каждая из которых распределены по закону $p(x|\theta)$. Предполагается, что функция $p(x|\theta)$ известная с точностью до параметра θ.

• Функция

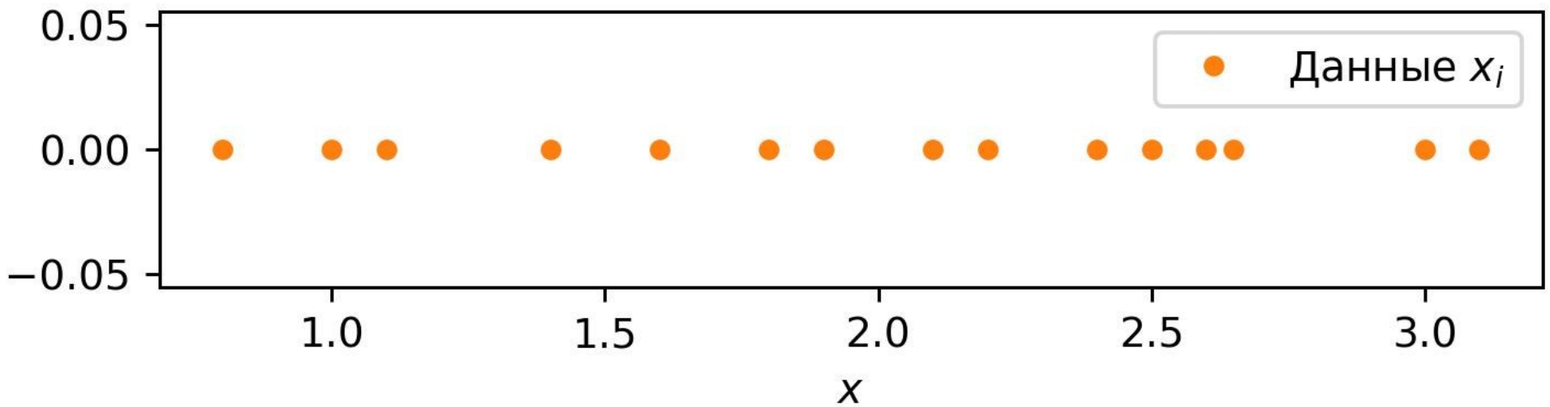
$$L(\theta) \coloneqq \prod_{i=1}^{n} p(x_i|\theta).$$

представляет собой случайную величину и называется ***функцией правдоподобия***.

• **Повторим**: метода ***максимального правдоподобия*** состоит в том, чтобы найти такое значение параметра θ, которое будет максимизировать $L(\theta)$. При максимизации подразумевается, что реализации с.в. $x_1, x_2, \dots, x_n$ фиксированы, а изменяется только параметр θ.

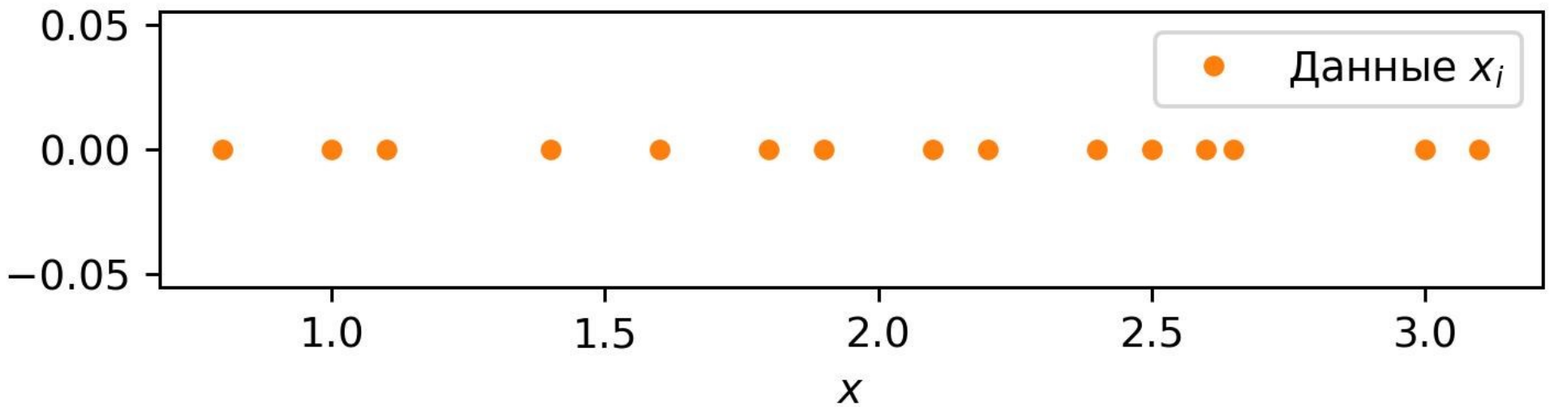
Максимальное правдоподобие: пример

• Пусть даны реализации с.в. $x_1, x_2, \ldots, x_n$ имеющей нормальное распределение и известное σ.



Максимальное правдоподобие: пример

• Пусть даны реализации с.в. $x_1, x_2, \dots, x_n$ имеющей нормальное распределение и известное σ.



• Функция правдоподобия имеет вид (пусть $\sigma = 0{,}8$):

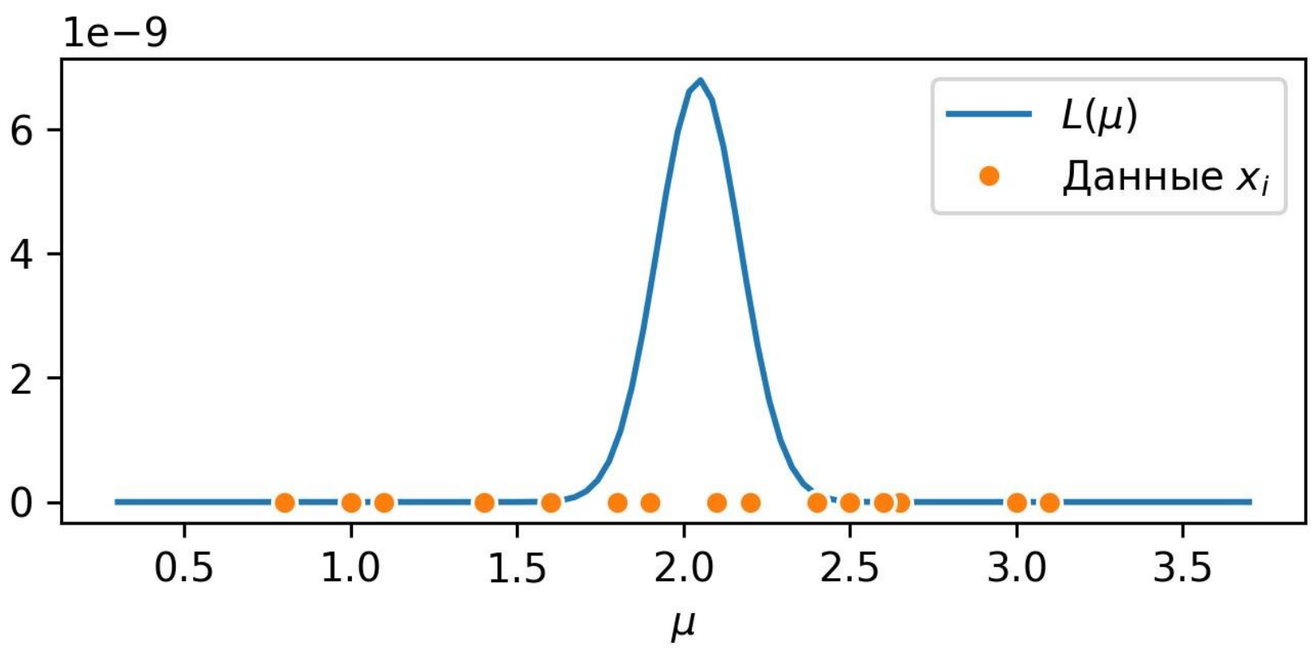
$$L(\mu) = \prod_{i=1}^{n} p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Максимальное правдоподобие: пример

• Пусть даны реализации с.в. $x_1, x_2, \dots, x_n$ имеющей нормальное распределение и известное σ.

• Функция правдоподобия имеет вид (пусть $\sigma = 0{,}8$):

$$L(\mu) = \prod_{i=1}^{n} p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

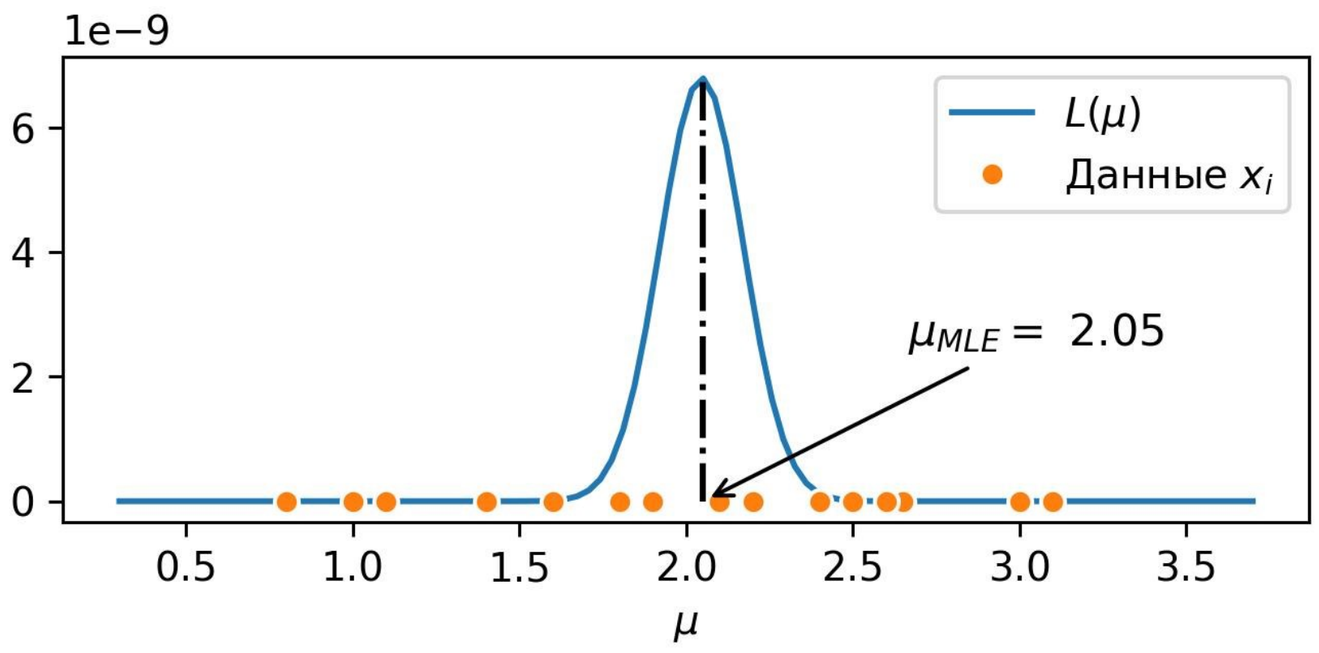


Максимальное правдоподобие: пример

• Пусть даны реализации с.в. $x_1, x_2, \ldots, x_n$ имеющей нормальное распределение и известное σ.

• Функция правдоподобия имеет вид (пусть $\sigma = 0{,}8$):

$$L(\mu) = \prod_{i=1}^{n} p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$



Максимальное правдоподобие: пример

- Функция правдоподобия имеет вид:

$$L(\mu) = \prod_{i=1}^{n} p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma^n}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2).$$

- На практике используют логарифм от функции правдоподобия:

$$\log L(\mu) = \log\prod_{i=1}^{n} p(x_i|\mu,\sigma) = \underbrace{\log\frac{1}{\sqrt{2\pi}\sigma^n}}_{\text{const}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

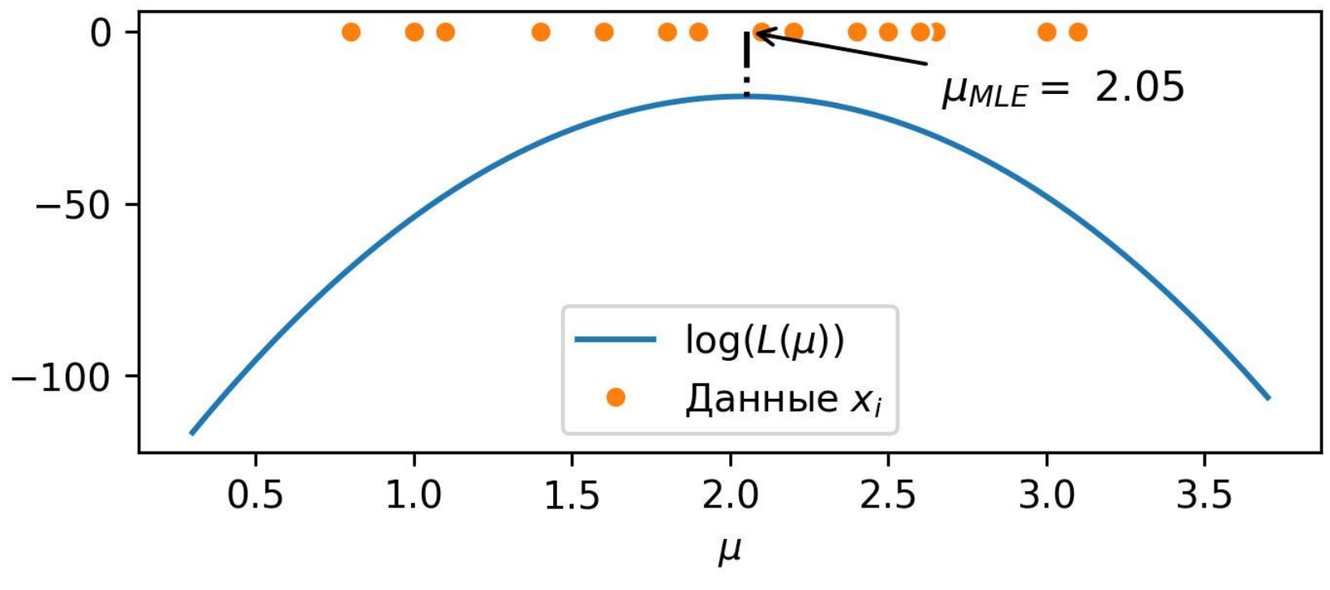
Максимальное правдоподобие: пример

- Функция правдоподобия имеет вид:

$$L(\mu) = \prod_{i=1}^{n} p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma^n}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2).$$

- На практике используют логарифм от функции правдоподобия:

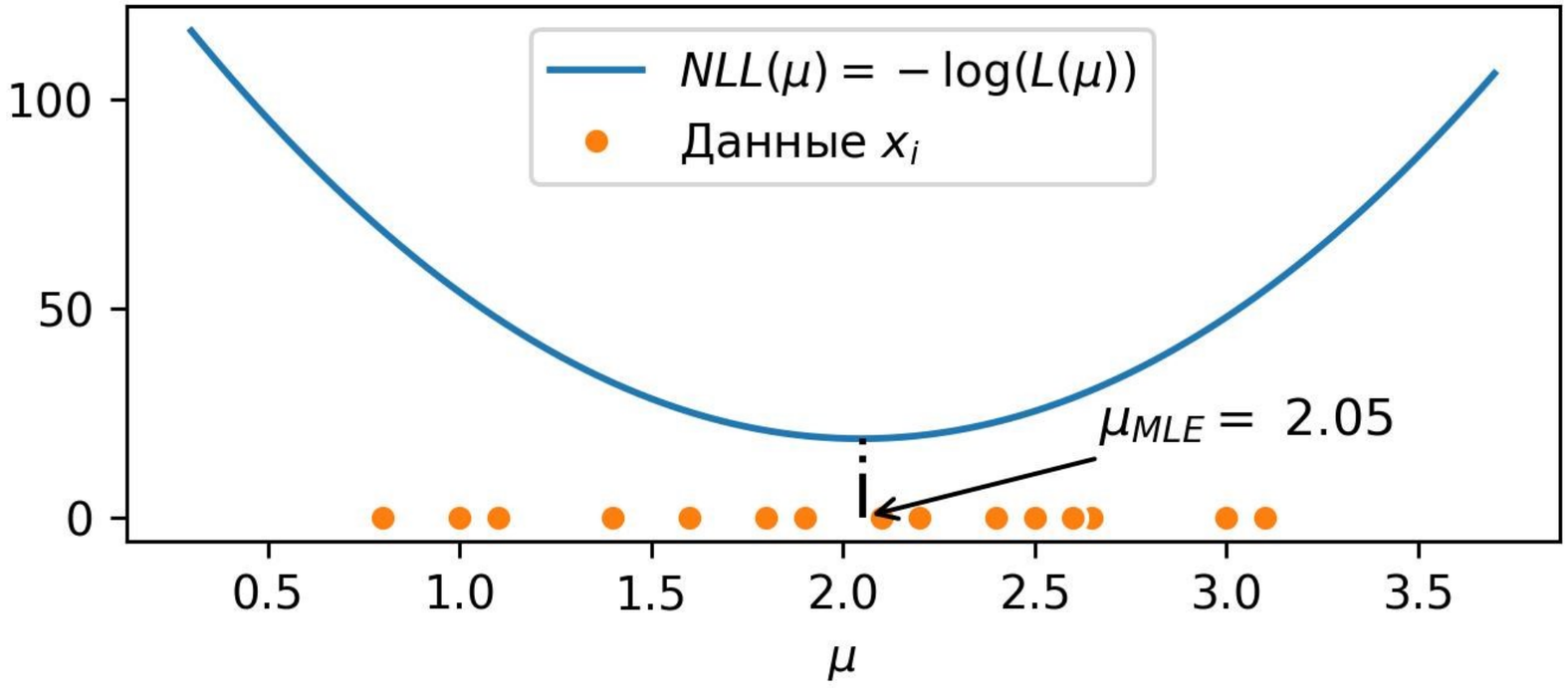
$$\log L(\mu) = \log\prod_{i=1}^{n} p(x_i|\mu,\sigma) = \underbrace{\log\frac{1}{\sqrt{2\pi}\sigma^n}}_{\text{const}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$



Максимальное правдоподобие: пример

• На практике также вместо поиска максимума функции правдоподобия ищут минимум отрицательного логарифмического правдоподобия:

$$NLL(\mu) = -\log L(\mu) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 + \text{const.}$$



Метод максимального правдоподобия для линейной регрессии

Правдоподобие

$$L_{\mathcal{D}}(\widehat{\boldsymbol{w}}) := \prod_{i=1}^{n} p(x_i, y_i) = \prod_{i=1}^{n} p(y_i|x_i)p(x_i) = \prod_{i=1}^{n} p(y_i|x_i) \prod_{i=1}^{n} p(x_i)$$

Метод максимального правдоподобия для линейной регрессии

Правдоподобие

$$L_{\mathcal{D}}(\widehat{\boldsymbol{w}}) \coloneqq \prod_{i=1}^{n} p(x_i, y_i) = \prod_{i=1}^{n} p(y_i|x_i)p(x_i) = \prod_{i=1}^{n} p(y_i|x_i) \prod_{i=1}^{n} p(x_i)$$

Условное правдоподобие

$$L_{\mathcal{D}}^{\text{cond}}(\widehat{\boldsymbol{w}}) \coloneqq \prod_{i=1}^{n} p(y_i|x_i) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\boldsymbol{x}_i^T\boldsymbol{w}, \hat{\sigma}^2\right) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{(y_i-\hat{y}_i)^2}{2\hat{\sigma}^2}} = \frac{1}{\left(\sqrt{2\pi}\hat{\sigma}\right)^n} e^{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

Метод максимального правдоподобия для линейной регрессии

Правдоподобие

$$L_{\mathcal{D}}(\widehat{\boldsymbol{w}}) \coloneqq \prod_{i=1}^{n} p(x_i, y_i) = \prod_{i=1}^{n} p(y_i|x_i)p(x_i) = \prod_{i=1}^{n} p(y_i|x_i) \prod_{i=1}^{n} p(x_i)$$

Условное правдоподобие

$$L_{\mathcal{D}}^{\text{cond}}(\widehat{\boldsymbol{w}}) \coloneqq \prod_{i=1}^{n} p(y_i|x_i) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\boldsymbol{x}_i^T\boldsymbol{w}, \hat{\sigma}^2\right) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{(y_i-\hat{y}_i)^2}{2\hat{\sigma}^2}} = \frac{1}{\left(\sqrt{2\pi}\hat{\sigma}\right)^n} e^{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

Логарифм правдоподобия

$$\log L_{\mathcal{D}}^{\text{cond}}(\widehat{\boldsymbol{w}}) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = -\frac{1}{2\hat{\sigma}^2}RSS - \text{const.}$$

$\Rightarrow$ если мы предполагаем нормальность ε, то оценка по методу максимального правдоподобия совпадает с оценкой по методу наименьших квадратов.

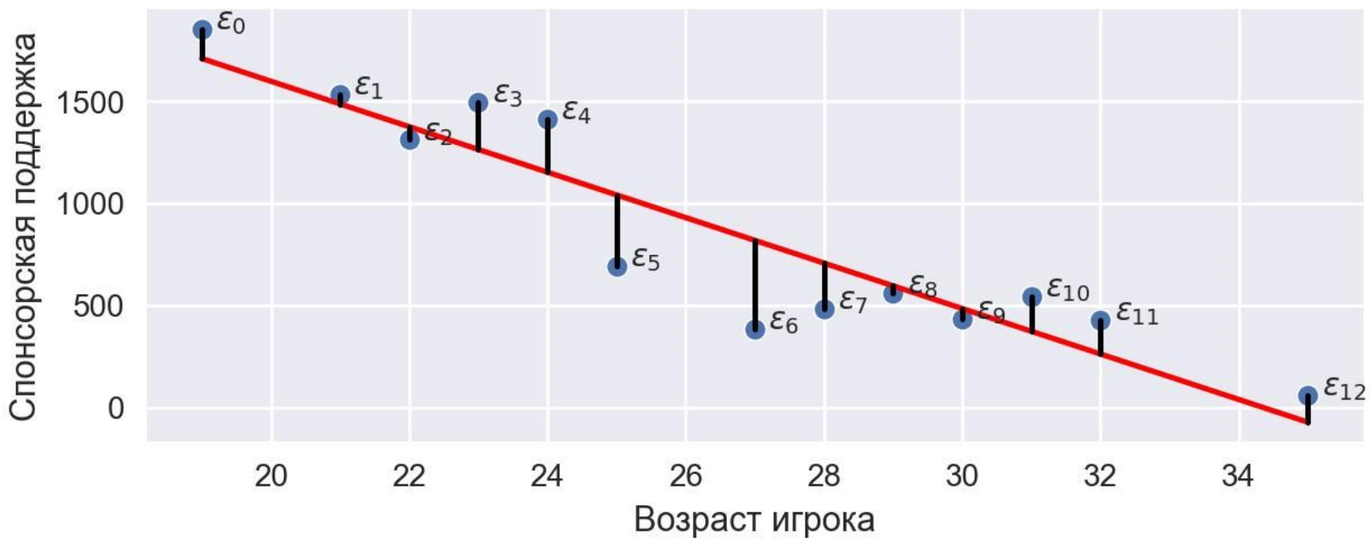
Оценка качества линейной регрессии

Значение RSS

Значение RSS показывает, какая величина отклонения (дисперсии) остается после подгонки линейной модели, которая измеряется квадратами различий между прогнозируемыми и фактическими целевыми значениями.

$$RSS = \sum_{i=1}^{n} {\varepsilon_i}^2$$

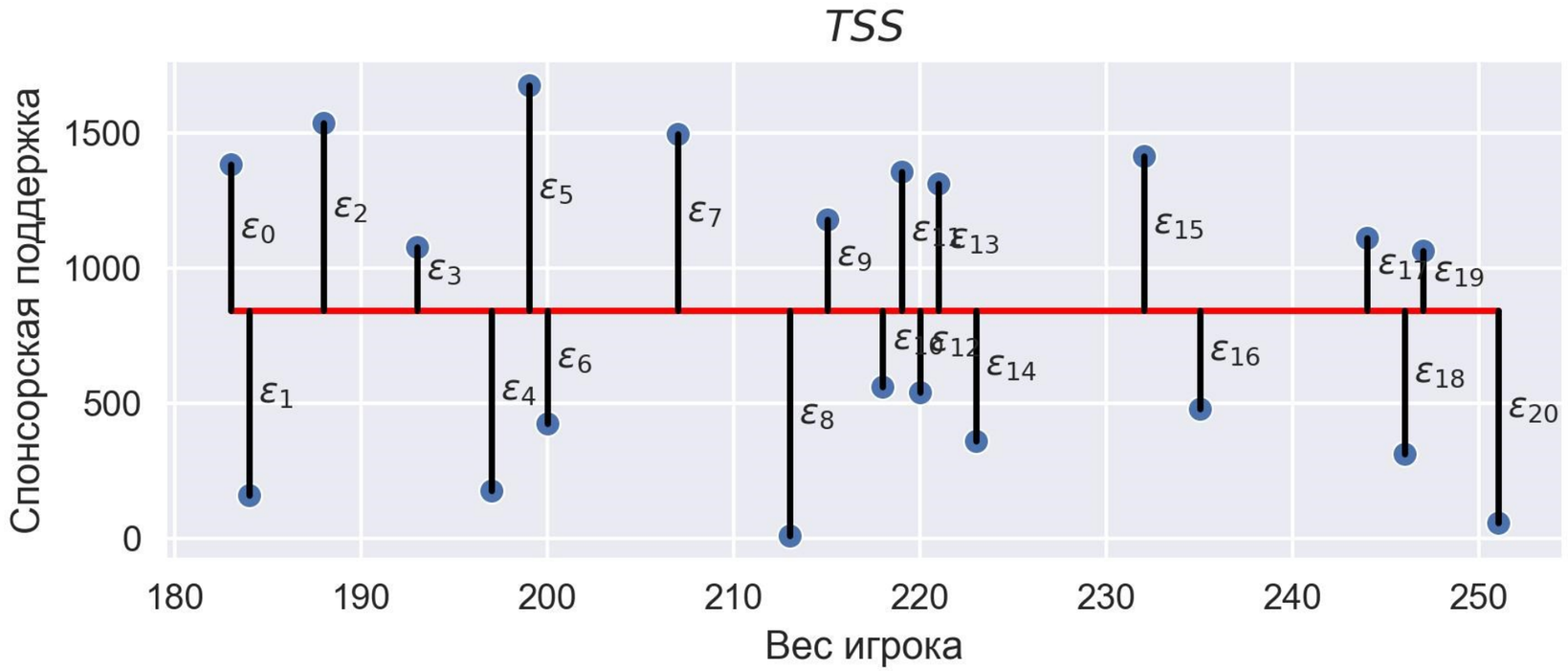
Недостаток RSS – его зависимость от набора данных и величин измерения целевой переменной.



Коэффициент детерминации R^2

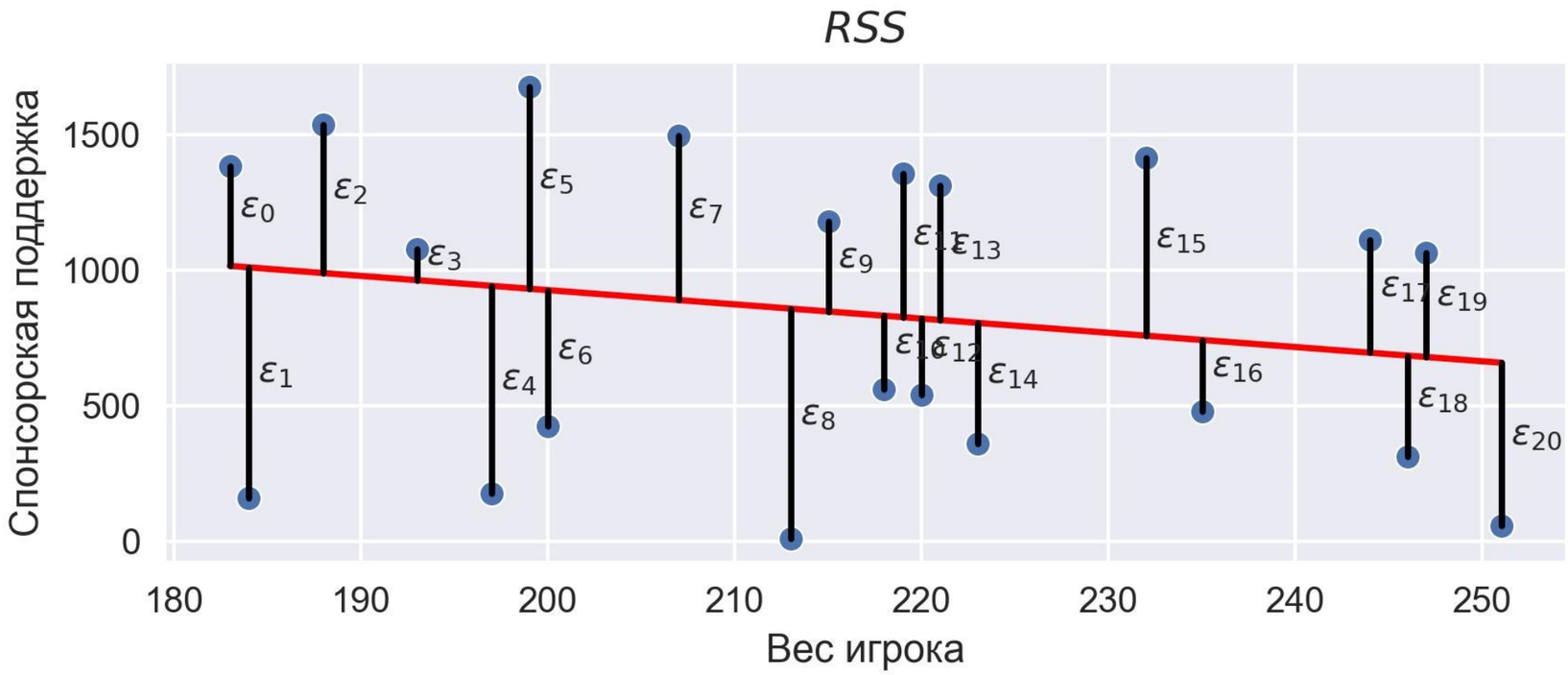
В основе расчета коэффициента R^2 лежит сравнение оцениваемой модели с «базовой» моделью. В качестве ***базовой модели*** используется модель, которая для любого x возвращает среднее значение целевой переменной y, которое рассчитано на всем наборе.

$$TSS = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$



Коэффициент детерминации R^2

$$R^2 = 1 - \frac{\text{сумма квадратов остатков}}{\text{общая сумма квадратов}} = 1 - \frac{RSS}{TSS},$$



$$R^2 = 1 - \frac{8\,778\,000}{9\,019\,000} = 0.026$$

Значение R^2 находится в диапазоне $[0, 1)$. Значения, близкие к единице указывают на лучшую эффективность модели. Коэффициент детерминации R^2 также интерпретируют, как величину изменения зависимой переменной y, которая обясняется независимыми (описательными) признаками x.

Скорректированный коэффициент R^2 (adjusted R^2)

R^2 всегда увеличивается (не уменьшается) с увеличением количества предикторов в модели, даже если они вообще не содержат никакой информации о целевой переменной.

Скорректированный коэффициент R^2 (adjusted R^2)

R^2 всегда увеличивается (не уменьшается) с увеличением количества предикторов в модели, даже если они вообще не содержат никакой информации о целевой переменной.

Пример

Допустим, что мы хотим построить линейную регрессию, чтобы предсказать размер спонсорской поддержки (y) по описательным признакам (Возраст, Рост, Вес). Рассчитаем несколько моделей каждый раз добавляя по одному новому признаку и посмотрим, как будет изменятся R^2. Результаты представлены в следующей таблице.

№	Предиктор	R^2
1	Возраст	0.8427
2	Рост	0.8629
3	Вес	0.8639

Скорректированный коэффициент R^2 (adjusted R^2)

Скорректированный $\boldsymbol{R^2}$ учитывает количество предикторов, используемых в модели:

$$R^2_{\text{adj}} = 1 - \left(\frac{RSS}{TSS} \times \frac{n-1}{n-p-1}\right),$$

где n – число наблюдений, p – число предикторов.

Пример (продолжение)

Если мы теперь посчитаем R^2_{adj} для данных о спонсорской поддержки, то получим следующие значения.

Таблица 0.1 – Коэффициент R^2_{adj}

№	Предиктор	$\boldsymbol{R^2_{\text{adj}}}$
1	Возраст	0.8427
2	Рост	0.8580
3	Вес	0.8539

Значимость коэффициентов регрессии

Найденные коэффициенты регрессии $\widehat{w}_j$ являются случайными величинами.

- Для проверки значимости коэффициентов регрессии выдвигается гипотеза

H_0: между X_j и Y нет взаимосвязи

Альтернативная гипотеза:

H_1: между X_j и Y есть взаимосвязь.

- Математически это выражается, как

$$H_0: w_j = 0$$

против

$$H_1: w_j \neq 0.$$

Значимость коэффициентов регрессии

Принятие/отклонение гипотезы выполняется в три этапа:

1) Вычисляется **тестовая статистика**.

2) Вычисляется вероятность того, что значение с.в. с распределением тестовой статистики будет больше или равно значению тестовой статистики. Эта вероятность называется p-значением.

3) P-значение сравнивается с предопределенным порогом значимости, и если p-значение меньше порога, то нулевая гипотеза отклоняется. Стандартные статистические пороги равны 5 и 1%.

Значимость коэффициентов регрессии

- Известно, что w_j распределено по нормальному закону.
- Дисперсия коэффициента w_j линейной регрессии рассчитывается как

$$\mathrm{SE}\left(\widehat{w}_j\right)^2 = \sigma_{w_j}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)},$$

- σ^2 мы не знаем, поэтому выборочная дисперсия служит её оценкой:

$$\sigma^2 \to s^2 = \frac{\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{w}}\right)^2}{n-2}.$$

- Статистика (предполагается, что $w_j = 0$)

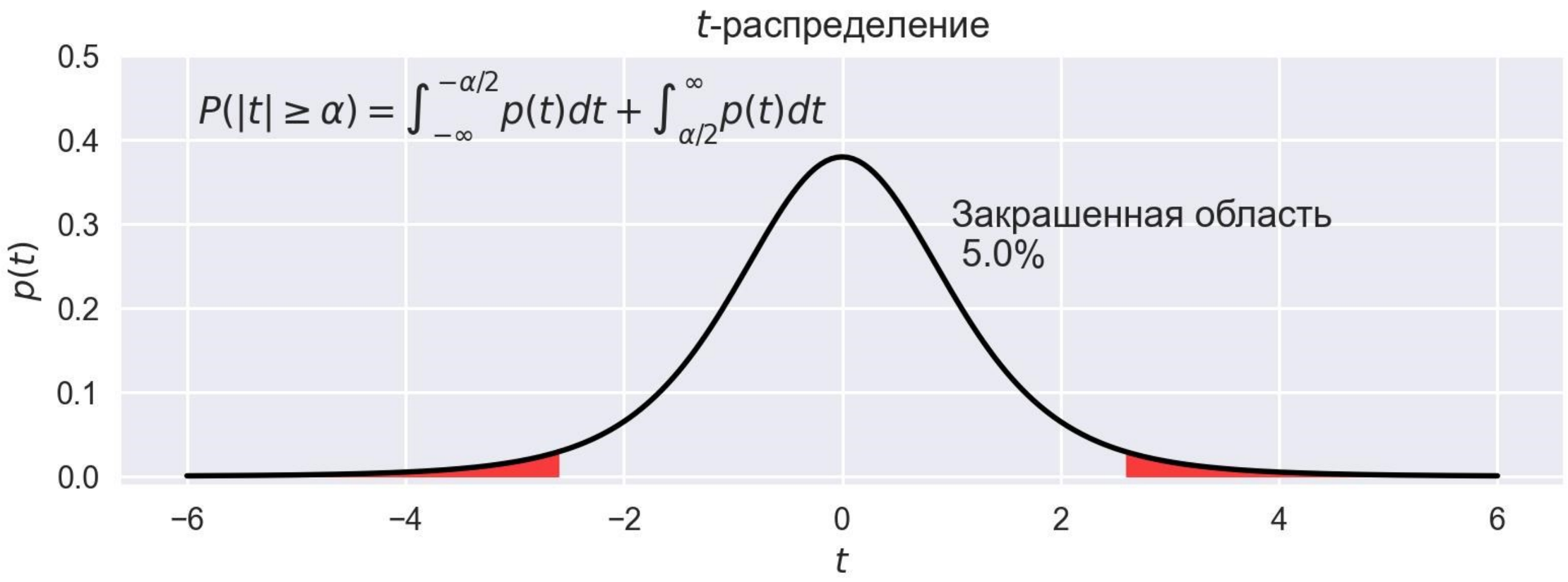
$$t = \frac{\widehat{w}_j - 0}{SE\left(\widehat{w}_j\right)},$$

подчинена t-распределению с $n - p - 1$ степенями свободы.

Значимость коэффициентов регрессии

$$t = \frac{\hat{w}_j - 0}{SE(\hat{w}_j)},$$

- Используя статистическое программное обеспечение, легко вычислить вероятность получения значения, равного или больше $|t|$. Эта вероятность называется ***p-значением***.



Значимость коэффициентов регрессии

Рассчитаем значимость коэффициентов регрессии для данных о баскетбольной команде.

№	Предиктор	Коэффициент w_j	Значение t	p-значение
1	Возраст	-128.76	-13.935	4.0515e-14
2	Рост	8.93	3.127	0.004
3	Вес	-2.08	-1.059	0.298

Доверительный интервал

С доверительной вероятностью 95% интервал

$$[\widehat{w}_j - 2 \times \mathrm{SE}(\widehat{w}_j), \quad \widehat{w}_j + 2 \times \mathrm{SE}(\widehat{w}_j)]$$

содержит реальное значение w_j (по сценарию, в котором мы получили повторяющиеся выборки, подобные имеющейся).

Интерпретация коэффициентов регрессии

- Идеальный случай – предикторы не коррелируют друг с другом:
  - Каждый коэффициент может быть оценен и протестирован отдельно.
  - Возможны такие интерпретации, как "*изменение в X_j связанном с w_j вызывает изменение Y, в то время как все остальные переменные остаются неизменными*".

- Наличие корреляции между предикторами создают проблемы:
  - Дисперсия всех коэффициентов имеет тенденцию к увеличению, иногда существенному
  - Интерпретации становятся опасными – когда меняется X_j, меняется все остальное.

- Следует избегать ***заявлений о причинно-следственной связи*** в отношении данных наблюдений.

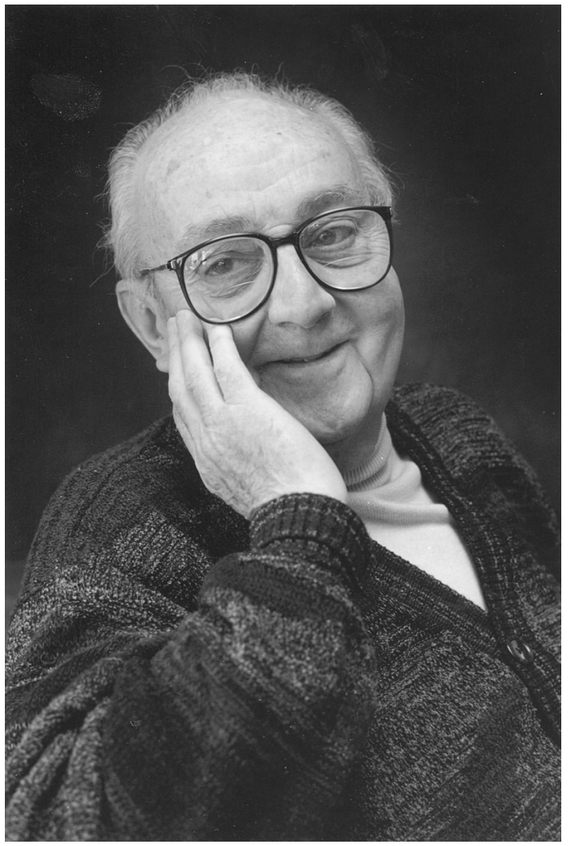
Интерпретация коэффициентов регрессии

• Коэффициент регрессии w_j оценивает ожидаемое изменение Y при изменении X_j на одну единицу, **при этом все остальные предикторы остаются неизменными**. Но предикторы обычно меняются одновременно!

Пример

Пусть Y - общая сумма мелочи в вашем кармане; X_1 = количество монет; X_2 = количество монет номиналом 1, 2, 5 и 10 копеек. Сам по себе коэффициент регрессии Y на X_2 будет > 0. А каким будет коэффициент при X_1?

Вместо заключения



В 1976 году британский статистик Джордж Бокс написал фразу: «***Все модели неправильны, но некоторые из них полезны***».

Он имел в виду, что мы должны сосредотачиваться на пользе моделей в прикладных сценариях, а не бесконечно спорить о том, является ли модель точной («правильной»)