Регрессия: предсказать число

Завтра +24°. Сколько прокатов ждать за день? Построим прогноз по пяти прошлым дням и проверим, насколько он ошибается.

Лето, вы подрабатываете в велопрокате у парка. Менеджер планирует завтрашнюю смену: если клиентов много, нужен помощник на выдаче. Прогноз обещает +24°. Сколько прокатов ждать за день? В журнале за прошлую неделю пять строк: температура и число прокатов. Попробуем получить прогноз из этих примеров. Число прокатов не равно нужному запасу велосипедов: один велосипед могут арендовать несколько раз.

В прошлом уроке ответом был класс: спам или нет. Здесь ответ — число, и «угадал / не угадал» уже не работает: прогноз 320 при факте 328 — отличный, прогноз 200 — плохой, хотя оба «не угадали». Значит, нужны две вещи: правило, которое из температуры делает число, и линейка, которой измеряют промах.

1. Вопрос «сколько?»

В 1.1 мы уже разделили два вопроса: «какой?» — классификация, и «сколько?» — **регрессия** (*regression*), ответ-число на шкале: сколько прокатов, через сколько минут приедет такси, сколько стоит квартира. Разница не в признаках, а в ответе: у класса нет «чуть-чуть мимо» (письмо либо спам, либо нет), а у числа есть — ошибка становится расстоянием, а не «да/нет». Сегодня мы впервые построим само правило, которое это число считает.

> Настоящее число прокатов — 250. Модель А предсказала 235, модель В — 300. Как правильно сравнить их ответы?
>
> - ○ Побеждает та модель, что дала целое число
> - ○ По расстоянию до факта: $|250-235| = 15$ у А, $|250-300| = 50$ у В — А точнее
> - ○ Обе ошиблись — значит, обе одинаково плохи
> - ○ Сравнивать нельзя: ни одна не угадала точно
> - ○ Побеждает та, что предсказала больше

Регрессия — это когда ответ можно промахнуть на чуть-чуть.

2. Прямая как модель

Допустим, журнал за пять дней выглядит так (температура днём и число прокатов):

День	Температура x, °C	Прокатов y
Пн	10	170
Вт	15	200
Ср	20	280
Чт	25	360

Ранее

В прошлый раз мы провели порог по одному признаку и получили классификатор: число → «спам / не спам». Сегодня признак тот же — число, но и ответ — число: не «какой класс», а «сколько». Пригодятся таблица «объект × признак» и привычка считать ошибки.

← Карта задач ИИ

← Данные и признаки: что машина знает о мире

← Классификация: как фильтр узнаёт спам

Пример

В лаборатории шага 4 — настоящие данные Capital Bikeshare: 731 день за 2011–2012 годы (UCI Machine Learning Repository). Там считаются поездки всей сети за сутки. Это другая величина, чем запас велосипедов в одном пункте: один велосипед может совершить несколько поездок.

История

Откуда слово. Фрэнсис Гальтон в 1886 году сравнивал рост родителей и детей и заметил: у очень высоких родителей дети в среднем ниже их, у очень низких — выше. Он назвал это «регрессией к посредственности». Слово прижилось, хотя сегодня регрессия — любое предсказание числа (Wikipedia).

День	Температура x, °C	Прокатов y
Пт	30	390

Теплее — больше прокатов, и рост похож на равномерный. Самое простое правило, которое это выражает, — прямая:

$$\hat{y} = kx + b$$

Крышка над y — знак «так говорит модель», в отличие от факта y. Число k — **наклон** (*slope*): на сколько прокатов вырастает прогноз при потеплении на один градус. Число b — **сдвиг** (*intercept*): прогноз при нуле градусов; в наших данных нет ни одного дня с такой температурой, поэтому этот прогноз пока не проверен. Сдвиг позволяет поднять или опустить всю прямую.

Возьмём $k = 12$, $b = 40$. Тогда при +24° модель говорит $40 + 12 \cdot 24 = 328$ прокатов. При +10° — 160 (факт был 170, близко). При +30° — 400 (факт 390).

> Модель проката: $\hat{y} = 8x + 59$, где x — температура в градусах, $\hat{y}$ — число прокатов за день. Прогноз погоды: +13°. Сколько прокатов предскажет модель?

Два числа, k и b, — и это вся модель.

3. Остаток — мера промаха

Сравним прогнозы на двух днях. У модели А промахи равны 10 и 30 прокатам, у модели Б — 20 и 0. Если складывать расстояния от прогноза до факта, получится 40 у А и 20 у Б: на этих двух днях точнее Б. Один удачный день ещё не делает модель лучшей — сравниваем все выбранные дни по одному правилу.

Знак промаха тоже полезен. При факте 170 и прогнозе 160 модель недооценила спрос на 10. При факте 200 и прогнозе 220 — переоценила на 20. Такие разности «факт минус прогноз» называют **остатками**. На графике это вертикальные отрезки между точками и прямой.

> За два дня было 170 и 200 прокатов. Модель А предсказала 160 и 230, модель Б — 150 и 200. Выбирают меньшую сумму расстояний между прогнозом и фактом. Какая модель точнее на этих днях?
>
> - ○ Модели одинаковы: каждая ошиблась хотя бы раз
> - ○ Б обязательно будет точнее на всех будущих днях
> - ○ А: её первый прогноз ближе к факту
> - ○ Б: сумма промахов 20, а у А — 40

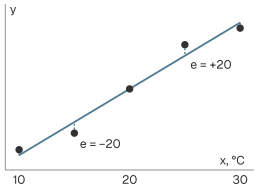
Профи. **Как посчитать MSE и RMSE**

Насколько прямая $k = 12$, $b = 40$ хороша? Для каждого дня посчитаем **остаток** (*residual*) — факт минус прогноз, $e = y - \hat{y}$:

Обозначения

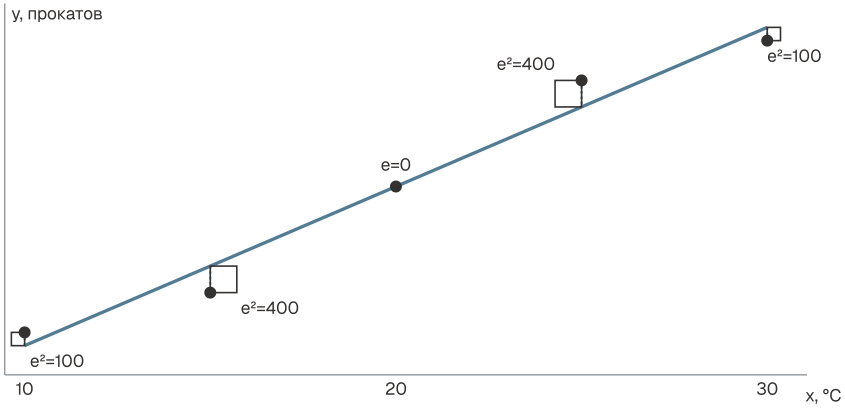
Символ	Смысл
x	температура, °C
y	прокатов по факту
$\hat{y}$	прогноз прокатов
k	наклон, прокатов на °C
b	прогноз при $x = 0$

Рисунок



Пять дней, прямая ŷ = 12x + 40 и остатки — вертикальные отрезки от точки до прямой. Среда лежит на прямой: остаток 0.

x	y	$\hat{y} = 12x + 40$	$e = y - \hat{y}$	e^2
10	170	160	+10	100
15	200	220	−20	400
20	280	280	0	0
25	360	340	+20	400
30	390	400	−10	100



Та же прямая, что и в таблице выше: на каждом остатке построен квадрат со стороной |e|. Площадь квадрата — это e^2, а сумма площадей — сумма квадратов остатков: 100 + 400 + 0 + 400 + 100 = 1000.

Сумма остатков равна нулю — но это не значит, что прямая идеальна: $+20$ и -20 просто погасили друг друга. В выводе МНК в шаге 5 станет ясно, что этот ноль не случаен.

Чтобы промахи не прятались за знаком, их возводят в квадрат — на рисунке выше это в буквальном смысле площадь. Сумма квадратов здесь $100 + 400 + 0 + 400 + 100 = 1000$, а её среднее — **среднеквадратичная ошибка** (*mean squared error, MSE*): $1000/5 = 200$. Корень из неё, $\sqrt{200} \approx 14$, — это **RMSE** (*root mean squared error*), величина ошибки в прокатах, тогда как сама MSE измеряется в прокатах в квадрате.

RMSE около 14 описывает ошибки на выбранных примерах. Это не гарантия, что ошибка прогноза на завтра окажется в пределах ±14.

Тот же расчёт кодом:

Остатки, MSE и RMSE на пяти днях

Цель: Запустите и сверьте напечатанные числа с таблицей остатков выше

```
import math

x = [10, 15, 20, 25, 30]          # температура, °C — признак
y = [170, 200, 280, 360, 390]     # прокатов — факт из журнала
k, b = 12, 40                     # прямая из шага 2: ŷ = 12x + 40

sse = 0                           # Σe² — сумма квадратов остатков
summa_e = 0                       # Σe — та же сумма, но без квадратов
print(f"{'x':>3} {'y':>6} {'ŷ':>6} {'e':>6} {'e²':>6}")
for xi, yi in zip(x, y):
    y_hat = k * xi + b            # прогноз модели на этот день
    e = yi - y_hat                # остаток: факт минус прогноз
    sse += e * e
    summa_e += e
    print(f"{xi:3d} {yi:6d} {y_hat:6d} {e:+6d} {e * e:6d}")

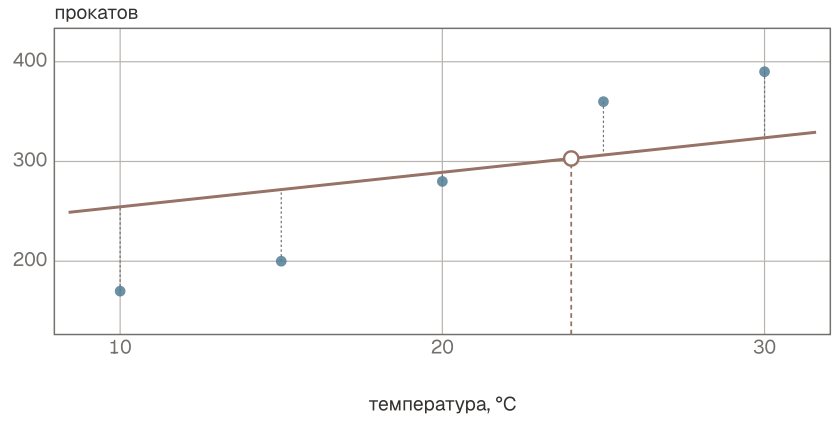
mse = sse / len(x)                # средний квадрат промаха
rmse = math.sqrt(mse)             # тот же промах, снова в прокатах
print(f"Σe = {summa_e}: плюсы погасили минусы — вот зачем нужен квадрат")
print(f"Σe² = {sse}")
print(f"MSE = {sse} / {len(x)} = {mse:.0f} (прокатов в квадрате)")
print(f"RMSE = √MSE ≈ {rmse:.1f} проката — величина ошибки")
print(f"Прогноз на завтра, +24°: {k * 24 + b} проката")
```

В опыте ниже используется другой показатель — **MSE**: среднее квадратов промахов. Он сильнее штрафует большие ошибки. Считать его вручную сейчас не нужно: меняйте линию и наблюдайте за числом. На тех же точках меньший показатель означает более точную подгонку.

Сначала запишите, какой прогноз даёт ваша линия на +24°. Затем попробуйте уменьшить MSE и сравните свой результат с кнопкой «Подобрать МНК»: она выбирает прямую с наименьшим значением этого показателя. Прогноз может измениться, хотя сами точки остались прежними.

Прямая, прогноз и цена ошибки

Пробная прямая и её ошибки



наклон k **3,46** сдвиг b **219,92** MSE **3 929**

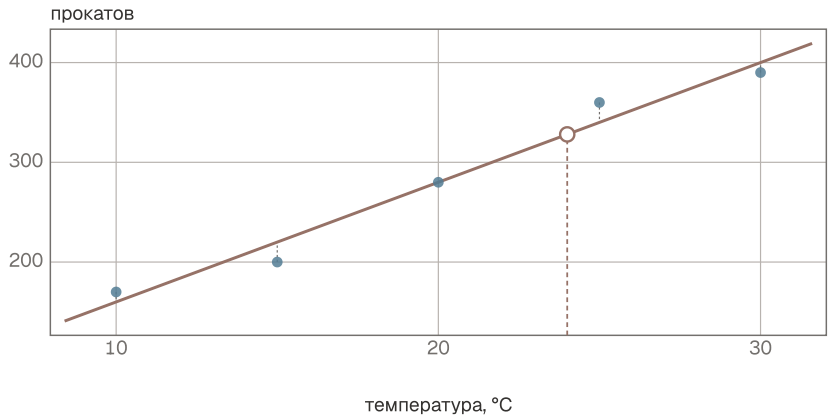
При 24,00 по оси «температура, °C» прогноз — 303 (прокатов). Это ответ модели, а не новое наблюдение.

▾

Заметка

Почему квадрат, а не модуль? Можно взять и модули: средняя абсолютная ошибка MAE здесь $(10+20+0+20+10)/5=1$. Но квадрат сильнее наказывает большие промахи (промах 20 «стоит» в 4 раза дороже промаха 10) и, главное, даёт формулу для лучшей прямой — шаг 5. А вот такой же короткой формулы через средние для модулей не существует — прямую с наименьшей суммой модулей ищут численно.

Прямая с наименьшей ошибкой

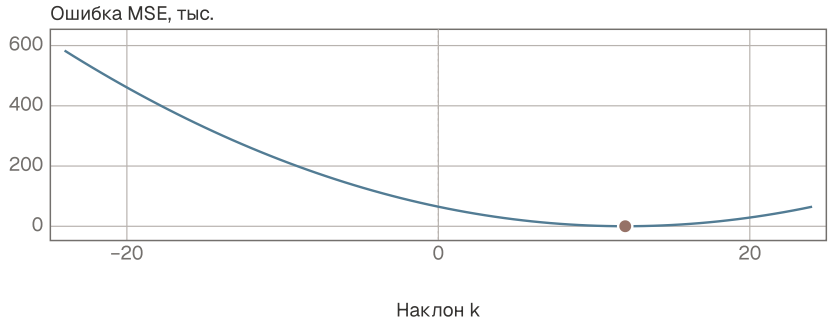


наклон k **12,00** сдвиг b **40,00** MSE **200**

При 24,00 по оси «температура, °C» прогноз — 328 (прокатов). Это ответ модели, а не новое наблюдение.

Как ошибка зависит от наклона

Каждая точка кривой — другая прямая: меняется только наклон, сдвиг b = 40,00 остаётся тем же. Отмечена ваша прямая; ниже — меньше ошибка.



наклон k **12,00** сдвиг b **40,00** MSE **200**

Тяните один конец прямой: второй остаётся на месте. Ползунок выбирает температуру для прогноза, пустой кружок показывает ответ модели. Ниже — ошибка при разных наклонах и текущем сдвиге. Попробуйте уменьшить MSE, затем сравните результат с кнопкой «Подобрать МНК».

Интерактивный эксперимент — в электронной версии урока.

Профи. Проверка расчёта ошибки

> За пять дней температура x и число прокатов y составили: (10°; 170), (15°; 200), (20°; 280), (25°; 360), (30°; 390). Модель прогнозирует $\hat{y} = 10x + 80$. Чему равна её MSE — среднее квадратов разностей между фактом и прогнозом?

На одном и том же наборе дней меньшая MSE означает более точную подгонку. Для выбора прогноза на будущее понадобится проверка на новых днях.

4. Эксперимент на настоящих данных

Пять дней были учебным примером. Теперь откроем все 731 наблюдение из файла `day.csv` набора Bike Sharing. Каждая точка — один день; по вертикали — поездки всей сети Capital Bikeshare за сутки. Температура записана в исходной шкале от 0 до 1: больше — теплее. Это не градусы Цельсия.

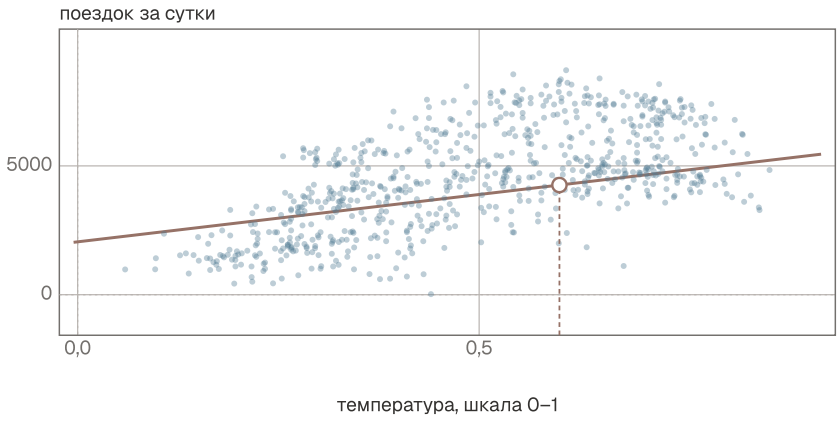
Перед подбором посмотрите на дни с похожей температурой. Одинаковое ли у них число поездок? Прямая даёт один прогноз, а наблюдения заметно различаются: одной температуры недостаточно, чтобы объяснить весь спрос.

Цитата

«Все модели неверны, но некоторые полезны.» — Джордж Бокс и Норман Дрейпер, *Empirical Model-Building and Response Surfaces*, 1987 (источник)

731 день настоящего велопроката

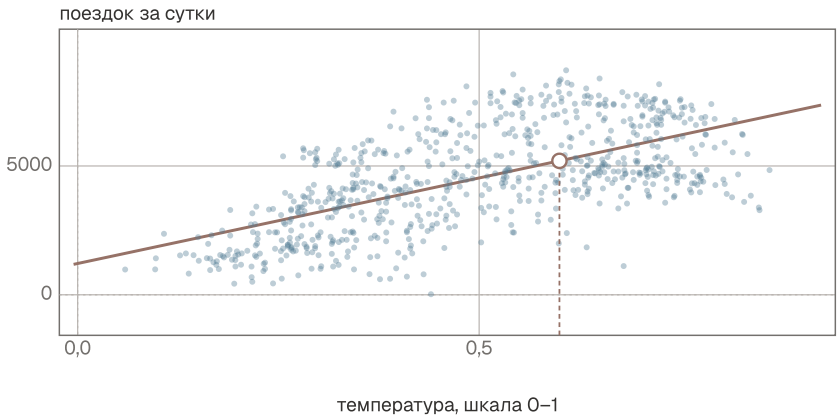
Пробная прямая и её ошибки



наклон k **3674,19** сдвиг b **2052,70** MSE **2 965 296**

При 0,60 по оси «температура, шкала 0–1» прогноз — 4 257 (поездок за сутки). Это ответ модели, а не новое наблюдение.

Прямая с наименьшей ошибкой



наклон k **6640,71** сдвиг b **1214,64** MSE **2 272 020**

При 0,60 по оси «температура, шкала 0–1» прогноз — 5 199 (поездок за сутки). Это ответ модели, а не новое наблюдение.

Сначала подберите прямую руками, затем нажмите «Подобрать МНК». Включите отрезки ошибок: даже лучшая прямая не попадёт во все точки. MSE измеряет подгонку по этим дням, а не качество прогноза на будущие дни.

Интерактивный эксперимент — в электронной версии урока.

Скачать исходные 731 наблюдение. Ползунок показывает ответ модели для выбранной температуры; нового измерения спроса при этом не возникает.

Профи. Что меняет сдвиг прямой

В прошлой лаборатории вы двигали и наклон, и сдвиг сразу. Разделим их: проверьте, что делает с остатками только сдвиг прямой вверх-вниз, отдельно от наклона.

Прогноз задан прямой $\hat{y} = kx + b$, остаток — $e = y - \hat{y}$ (факт минус прогноз). Сдвиг b увеличили на 20, не меняя наклона и наблюдений. Что произойдёт с остатками?

- ○ Остатки не изменятся, изменится только MSE
- ○ Сумма квадратов остатков обязательно уменьшится
- ○ Каждый остаток уменьшится ровно на 20 — и может стать отрицательным
- ○ Каждый остаток увеличится ровно на 20
- ○ Остатки изменятся по-разному: у крайних точек сильнее, чем у средних

Подумай

Подумай. Прямая из учебного примера, $k = 12$, $b = 40$, при +45° предскажет 580 прокатов.
В пятидневном журнале таких температур нет. Верите ли вы прогнозу за пределами того, что модель видела? Что физически происходит с желанием кататься при +45°?

В журнале при одной температуре было 300 и 500 поездок. Модель использует только температуру и всегда даёт одному входу один ответ. Что следует из этого?

- ○ Одним прогнозом нельзя точно попасть в оба значения спроса
- ○ В обоих днях факт обязательно записан с ошибкой
- ○ Нужно провести прямую через каждую точку одновременно
- ○ Температура вообще не может быть полезным признаком

Реальное облако шире: полезный прогноз может ошибаться, а точность нужно измерять.

5. Проверка на новых днях

Кнопка подбора уменьшает ошибку на уже известных точках. Но завтра появится новая точка. Чтобы проверить пользу модели, запишем прогнозы заранее, дождёмся факта и сравним их на днях, по которым линию не настраивали.

Представим две модели проката. На старых днях А ошибалась меньше. Для двух следующих дней прогнозы зафиксировали заранее:

Новый день	Факт	Прогноз А	Прогноз Б
Суббота	300	290	280
Воскресенье	360	310	350

Расстояния до факта: у А — 10 и 50, у Б — 20 и 10. По сумме этих расстояний на новых днях выигрывает Б: 30 против 60. На меньшей ошибке по старому журналу выбор не заканчивается.

На двух новых днях промахи модели А составили 10 и 50 прокатов, модели Б — 20 и 10. На старых днях, использованных для настройки, А была точнее. Какой вывод подтверждает новая проверка по сумме промахов?

- ○ Ни одну модель нельзя применять, пока все промахи не равны нулю
- ○ Б теперь всегда будет точнее А
- ○ На этих двух новых днях Б точнее, но двух дней мало для гарантии на будущее
- ○ А лучше: ошибка на старых днях важнее любой новой проверки

Если после просмотра этой таблицы поправить модель, таблица станет частью настройки. Для следующей независимой проверки нужны другие дни. Это полезная привычка и без формул: сначала прогноз, потом факт, затем честный разбор ошибки.

Есть и границы применения. В учебном журнале температура была от +10° до +30°. При +45° формула всё ещё выдаст число, но наблюдений для проверки такого режима у нас нет. Перед использованием спросите: похожи ли новые условия на те, по которым модель строилась?

Профи. Формула метода наименьших квадратов

Подбирать прямую вручную — долго, и нет гарантии, что нашли лучшую. К счастью, у задачи «минимальная сумма квадратов остатков» есть точный ответ — **метод наименьших квадратов** (*least squares*, МНК). Нужны средние $\bar{x}$ и $\bar{y}$ и отклонения от них:

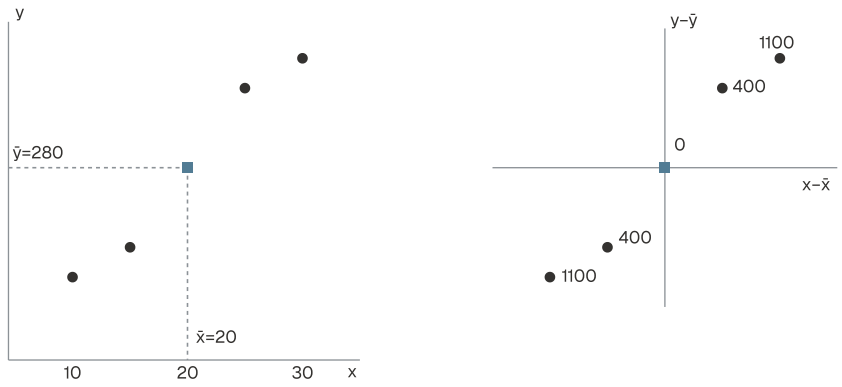
$$\boxed{k = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad b = \bar{y} - k\,\bar{x}}$$

Почему именно эта пара k, b даёт наименьшую сумму квадратов среди всех прямых — докажем в следующем блоке; пока проверим расчёт на пяти днях. И ещё: делить на $\sum (x_i - \bar{x})^2$ можно, только если температуры не все одинаковы — иначе знаменатель равен нулю, а по одной температуре наклон не определить.

Здесь $\bar{x}$ — средняя температура, $\bar{y}$ — средний спрос, x_i, y_i — i-й день, $\sum$ — сумма по всем дням. Числитель показывает, насколько x и y отклоняются от средних *вместе*; знаменатель — насколько разбросан сам x.

Пересчитаем для пяти дней. $\bar{x} = (10 + 15 + 20 + 25 + 30)/5 = 20$, $\bar{y} = 1400/5 = 280$.

$x_i - \bar{x}$	$y_i - \bar{y}$	**произведение**	$(x_i - \bar{x})^2$
−10	−110	1100	100
−5	−80	400	25
0	0	0	0
5	80	400	25
10	110	1100	100
Сумма	0	**3000**	**250**



Слева — пять дней в осях (x, y), пунктир — точка средних (x̄, ȳ) = (20, 280). Справа — те же точки в осях отклонений (x−x̄, y−ȳ): подписаны произведения (xᵢ−x̄)(yᵢ−ȳ) из таблицы выше. Наклон k = 3000/250 — это то, насколько облако «повёрнуто» вокруг своего центра.

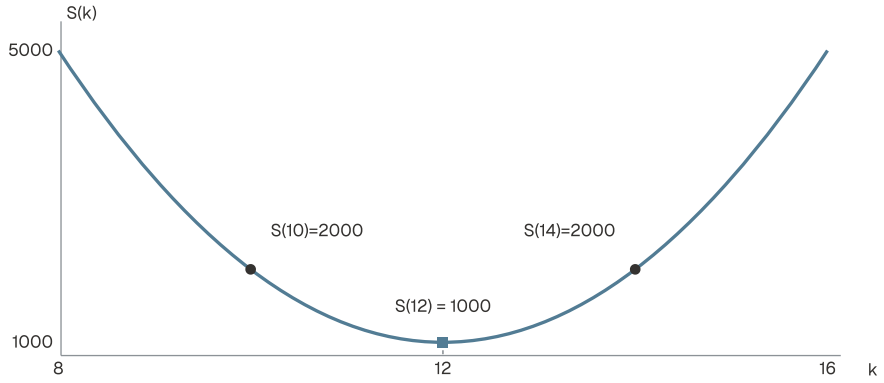
$k = 3000/250 = 12$, $b = 280 - 12 \cdot 20 = 40$. Прямая из шага 2 и есть прямая наименьших квадратов. Формула даёт MSE = 200; доказательство ниже устанавливает, что меньшей MSE на этих пяти днях у прямой со свободным сдвигом быть не может.

▾ *Обозначения*

Символ	Смысл
$\bar{x}, \bar{y}$	средние по всем n дням
$\sum$	сумма по все дням
n	число дней
e^2	квадрат оста
MSE	$\sum e^2/n$ — средний квадрат промаха
S_{xx}, S_{xy}	$\sum (x_i - \bar{x})^2$ и $\sum (x_i - \bar{x})(y_i -$

Заметьте: $b = \bar{y} - k\bar{x}$ означает, что при $x = \bar{x}$ прогноз равен $\bar{y}$. Прямая МНК всегда проходит через точку средних $(\bar{x}, \bar{y})$ — у нас через $(20, 280)$, и это среда, где остаток 0.

> У другого проката средняя температура $\bar{x} = 18$, средний спрос $\bar{y} = 381$, а наклон МНК уже посчитан: $k = 11$. Найдите сдвиг b.



Сумма квадратов остатков S(k) при подставленном b = ȳ − k·x̄ — парабола по k. Вершина в k = 12 даёт минимум S = 1000; в k = 10 и k = 14 сумма одинаково хуже — 2000. «Минимум» — это буквально дно этой параболы.

> У другого набора дней $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = 132$, а $S_{xx} = \sum(x_i - \bar{x})^2 = 46$. Чему равен наклон k прямой наименьших квадратов?

Теперь напишите формулу сами:

Формула наименьших квадратов своими руками

Цель: Допишите две строки в mnk, чтобы функция считала наклон и сдвиг МНК на данных с разными значениями признака

```
x = [10, 15, 20, 25, 30]          # те же пять дней из журнала
y = [170, 200, 280, 360, 390]

def mnk(x, y):
    """Наклон k и сдвиг b прямой наименьших квадратов (формула шага
5)."""
    n = len(x)
    x_sr = sum(x) / n             # x̄ — средняя температура
    y_sr = sum(y) / n             # ȳ — средний спрос
    s_xy = 0.0                    # S_xy = Σ(xᵢ−x̄)(yᵢ−ȳ) — числитель
    s_xx = 0.0                    # S_xx = Σ(xᵢ−x̄)² — знаменатель
    for xi, yi in zip(x, y):
        s_xy += 0.0               # TODO: прибавьте произведение
отклонений от средних
        s_xx += 1.0               # TODO: прибавьте квадрат отклонения по
x
    k = s_xy / s_xx
    b = y_sr - k * x_sr           # прямая проходит через точку средних
    return k, b

k, b = mnk(x, y)
mse = sum((yi - (k * xi + b)) ** 2 for xi, yi in zip(x, y)) / len(x)
print(f"k = {k:.2f} прокатов на градус, b = {b:.2f}")
print(f"MSE = {mse:.0f}; у прямой ŷ = 12x + 40 из шага 3 было 200")
print(f"Прогноз при +24°: {k * 24 + b:.0f} проката")
```

Профи. Вывод формулы МНК без производных: две параболы

Сумма квадратов остатков как функция двух чисел k и b:

▾

Ссылки

- ↗ scikit-learn: линейные модели — как эта же формула выглядит в три строки кода (`LinearRegression`).
- ↗ Учебник ML Яндекса, «Линейные модели» — тот же МНК, но сразу для многих признаков.
- ↗ StatQuest — короткие видео про регрессию и R^2, на пальцах.

$$S(k, b) = \sum_{i=1}^{n} (y_i - k x_i - b)^2.$$

Шаг 1: фиксируем k, ищем лучший b. Раскроем квадраты и соберём по степеням b:

$$S = n b^2 - 2b \sum (y_i - k x_i) + \text{const}.$$

Это парабола по b с ветвями вверх ($n > 0$); её вершина — в точке $b = \frac{1}{n}\sum (y_i - k x_i) = \bar{y} - k\bar{x}$. Это и есть вторая формула.

Для любых k и b верно $S(k, b) \geq S(k, b^*(k))$, где $b^*(k)$ — вершина этой параболы по b; значит минимум S по обеим переменным равен минимуму $S(k) = S(k, b^*(k))$ по одному k — его и ищем дальше.

Шаг 2: подставляем и ищем k. С таким b остаток переписывается через отклонения: $y_i - k x_i - b = (y_i - \bar{y}) - k(x_i - \bar{x})$. Обозначим $S_{xx} = \sum (x_i - \bar{x})^2$, $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \sum (y_i - \bar{y})^2$. Тогда

$$S(k) = S_{yy} - 2k\, S_{xy} + k^2 S_{xx}$$

— снова парабола, теперь по k. При $S_{xx} > 0$ её вершина находится в $k = S_{xy} / S_{xx}$. Первая формула доказана.

Если все x_i одинаковы, то $S_{xx} = S_{xy} = 0$: после выбора $b = \bar{y} - k\bar{x}$ значение $S(k) = S_{yy}$ постоянно. Минимум достигается при любом k с соответствующим b; данные не определяют единственный наклон.

Бонус: чему равен минимум. Подставив вершину, получаем $S_{\min} = S_{yy} - S_{xy}^2 / S_{xx}$. Для наших пяти дней: $S_{yy} = 110^2 + 80^2 + 0 + 80^2 + 110^2 = 37000$, $S_{xy} = 3000$, $S_{xx} = 250$: $S_{\min} = 37000 - 9\,000\,000 / 250 = 1000$. Ровно то, что мы насчитали в таблице — теперь с гарантией, что меньше нельзя.

Доля объяснённого разброса $R^2 = 1 - S_{\min}/S_{yy} = 1 - 1000/37000 \approx 0{,}973$: температура «объясняет» 97 % разброса спроса в учебном примере. На реальных данных одной температуры недостаточно: дни с похожей температурой имеют разный спрос. Проверить вклад дождя или выходных можно, добавив эти признаки и сравнив качество на новых днях; по одному облаку точек причину разброса установить нельзя.

Все значения y умножили на 2 (считаем прокаты не в штуках, а в «полуштуках»). Что случится с прямой МНК?

- ○ Удвоится только k, сдвиг сохранится
- ○ k и b удвоятся, остатки удвоятся, R^2 не изменится
- ○ Ничего не изменится: МНК не зависит от масштаба
- ○ Удвоится только b, наклон сохранится
- ○ R^2 удвоится

Эпик. **Четыре свойства прямой наименьших квадратов**

Пусть $\bar{x}$, $\bar{y}$ — средние, $S_{xx} = \sum_i (x_i - \bar{x})^2 > 0$, $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \sum_i (y_i - \bar{y})^2 > 0$. Прямая МНК имеет коэффициенты $k = S_{xy}/S_{xx}$, $b = \bar{y} - k\bar{x}$, прогнозы $\hat{y}_i = kx_i + b$ и остатки $e_i = y_i - \hat{y}_i$. Доля объяснённого разброса: $R^2 = 1 - \sum_i e_i^2 / S_{yy}$. Докажите:

1. $\sum e_i = 0$ — остатки в сумме дают ноль.
2. $\sum x_i e_i = 0$ — остатки «не коррелируют» с признаком.
3. $S_{yy} = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2$ — полный разброс распадается на объяснённый и необъяснённый. Отсюда $0 \le R^2 \le 1$.
4. Если построить регрессию наоборот — x по y, с наклоном k', — то $k \cdot k' = R^2$.

Затем посчитайте R^2 двумя способами (через п. 3 и через п. 4) для пяти пар «температура; число прокатов»: (10°; 170), (15°; 200), (20°; 280), (25°; 360), (30°; 390).

▾*Решение*

1. $\sum e_i = \sum y_i - k \sum x_i - nb = n\bar{y} - nk\bar{x} - n(\bar{y} - k\bar{x}) = 0$.

2. Подставим $b = \bar{y} - k\bar{x}$: $e_i = (y_i - \bar{y}) - k(x_i - \bar{x})$. Тогда $\sum x_i e_i = \sum x_i (y_i - \bar{y}) - k \sum x_i (x_i - \bar{x})$. Заметим, что $\sum \bar{x}(y_i - \bar{y}) = 0$ и $\sum \bar{x}(x_i - \bar{x}) = 0$, поэтому $\sum x_i (y_i - \bar{y}) = S_{xy}$ и $\sum x_i (x_i - \bar{x}) = S_{xx}$. Получаем $S_{xy} - kS_{xx} = 0$ по определению k.

3. $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$. Возведём в квадрат и просуммируем; перекрёстный член $2\sum (\hat{y}_i - \bar{y}) e_i = 2\sum (kx_i + b - \bar{y}) e_i = 2k \sum x_i e_i + 2(b - \bar{y}) \sum e_i = 0$ по пп. 1–2. Значит $S_{yy} = \mathrm{SSR} + \mathrm{SSE}$, где $\mathrm{SSR} = \sum (\hat{y}_i - \bar{y})^2 \ge 0$ и $\mathrm{SSE} = \sum e_i^2 \ge 0$. Поэтому $R^2 = 1 - \mathrm{SSE}/S_{yy} = \mathrm{SSR}/S_{yy} \in [0, 1]$.

4. $k = S_{xy}/S_{xx}$, по симметрии $k' = S_{xy}/S_{yy}$, произведение $S_{xy}^2 / (S_{xx} S_{yy})$. С другой стороны, $\hat{y}_i - \bar{y} = k(x_i - \bar{x})$, поэтому $\mathrm{SSR} = k^2 S_{xx} = S_{xy}^2 / S_{xx}$ и $R^2 = \mathrm{SSR}/S_{yy} = S_{xy}^2/(S_{xx}S_{yy})$. Совпало.

Числа. $S_{xx} = 250$, $S_{xy} = 3000$, $S_{yy} = 37000$. Через п. 3: $\mathrm{SSR} = 144 \cdot 250 = 36000$, $R^2 = 36000/37000 \approx 0{,}973$. Через п. 4: $k' = 3000/37000 \approx 0{,}0811$, $k \cdot k' = 12 \cdot 0{,}0811 \approx 0{,}973$. Заодно видно, почему две регрессии — y по x и x по y — это разные прямые: они совпадают только при $R^2 = 1$.

> Для прямой МНК со свободным сдвигом известны суммы $S_{xx} = \sum_i (x_i - \bar{x})^2 = 45$, $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = 74$, $S_{yy} = \sum_i (y_i - \bar{y})^2 = 431$. Здесь $\bar{x}$ и $\bar{y}$ — средние. Найдите R^2, долю объяснённого разброса, с точностью до сотых.

Подбор линии — только начало. Полезность прогноза проверяют по наблюдениям, которые не участвовали в настройке.

Где это встретится дальше

Сумма квадратов остатков — пример **функции потерь**: по ней сравнивают настройки модели. В уроке «Обучение нейронных сетей: функция потерь» этот вопрос разбирается на других примерах. После первых четырёх уроков можно перейти к уроку «Ближайшие соседи: чей это цветок?»: другой способ прогноза опирается на похожие наблюдения. Для любого способа остаётся одна проверка: как он работает на данных, по которым его не настраивали.

Построили прогноз правило «умножить температуру на 12 и прибавить 40» даёт 328 прокатов при +24°. Это ответ модели, а не известный факт.

Сравнили ошибки расстояния от прогнозов до наблюдений позволяют сопоставить модели. Сравнивать нужно на одних данных и по одному правилу.

Проверили границу вывода хорошая подгонка старого журнала не гарантирует точности завтра. Нужны новые дни и проверка условий применения.

Домашняя работа

1. *Профи.* **Самостоятельно подобрать прямую**

   > Журнал проката за четыре дня: $x = 1, 2, 3, 4$ часа солнца, соответствующий спрос $y = 3, 5, 8, 10$ прокатов. Найдите наклон k прямой МНК $\hat{y} = kx + b$ со свободным сдвигом. Ответ с точностью до сотых.

2. Модель цены квартиры: $\hat{y} = 150\,S + 1200$, где $\hat{y}$ — цена в тысячах рублей, S — площадь в м². Что означает коэффициент 150?

   Каждый дополнительный квадратный метр добавляет к прогнозу 150 тысяч

   Квартира площадью 1 м² стоит 150 тысяч

   Модель ошибается в среднем на 150 тысяч

   Прогноз для несуществующей квартиры площадью 0 м²

   150 квартир участвовало в обучении

3. На двух новых днях столовая продала 20 и 30 порций. Модель А заранее предсказала 18 и 26, модель Б — 15 и 30. На бумаге отметьте факт и оба прогноза; выберите модель по сумме расстояний до факта. Объясните, почему одного точного попадания недостаточно.
4. Для прогноза поездок предложите один признак кроме температуры. Объясните, как его записать в таблицу и какие новые дни понадобятся, чтобы проверить, стало ли лучше. Направление влияния пока считайте предположением.

Источники

- D2L: Linear Regression (https://d2l.ai/chapter_linear-regression/linear-regression.html): линейный прогноз, квадратичная ошибка и подбор параметров.
- Bike Sharing Dataset — UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset) — данные Capital Bikeshare за 2011–2012: 731 день, погода и число прокатов.
- Simple linear regression — Wikipedia (https://en.wikipedia.org/wiki/Simple_linear_regression) — формулы МНК для одного признака, свойства остатков, R^2.
- Least squares — Wikipedia (https://en.wikipedia.org/wiki/Least_squares) — история: Лежандр, Гаусс и Церера.
- scikit-learn: Linear Models (https://scikit-learn.org/stable/modules/linear_model.html) — та же регрессия в коде.
- Учебник по машинному обучению Яндекса: линейные модели (https://education.yandex.ru/handbook/ml/article/linear-models) — МНК для многих признаков, регуляризация.
- Anscombe's quartet — Wikipedia (https://en.wikipedia.org/wiki/Anscombe%27s_quartet) — четыре разных облака точек с одинаковой прямой МНК: почему нужно смотреть на картинку.