# Регуляризация линейной регрессии

Маргарита Широбокова

Product Owner R&D ELMA AI

Skillbox образовательная платформа

Регуляризация линейной регрессии

# Анализ результатов линейной регрессии

# Уравнение регрессии

$$\hat{y} = f(w, x) = w_0 + w_1 \times x_1 + \ldots + w_k \times x_k$$

Где

$\hat{y}$ — целевая переменная,

$(x_1, \ldots, x_k)$ — вектор признаков,

$w_0, w_1, \ldots, w_k$ — параметры модели цели,

как их ещё называют

$w_1, \ldots, w_k$ — вектор весов,

а число $w_0$ — свободный коэффициент, или сдвиг (bias)

Или компактная запись

$$\hat{y} = \langle x, w \rangle + w_0$$

# Остатки регрессии

Сравнить предсказанные значения с реальными данными:

то есть сравнить y– и y



$$e = y – y –$$

y — реальное значение;

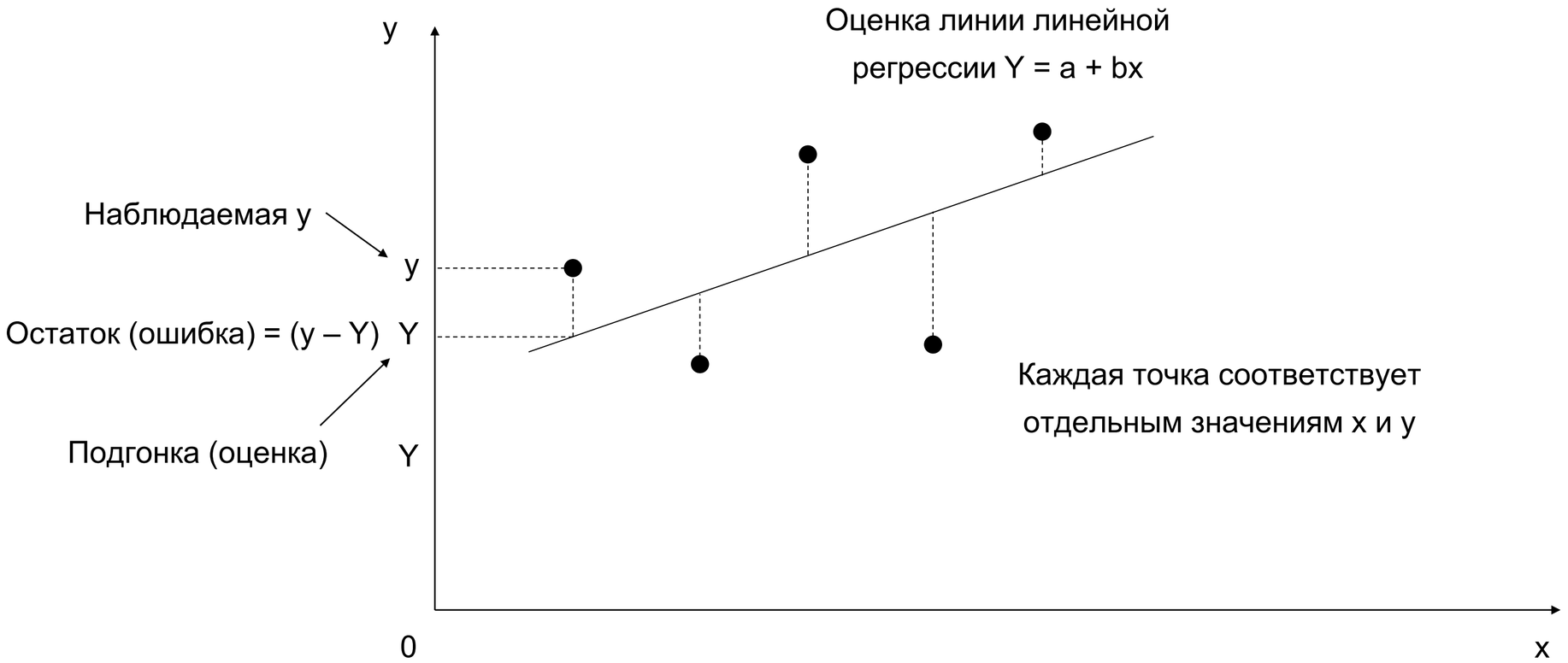
y– — предсказанное значение.

Остатки регрессии — это разности между реальными значениями и значениями, предсказанными регрессионной моделью.

В терминах матриц можно записать также:

$$e = y – y – = y – X\,(XTX) – 1XTy$$

# Остатки регрессии



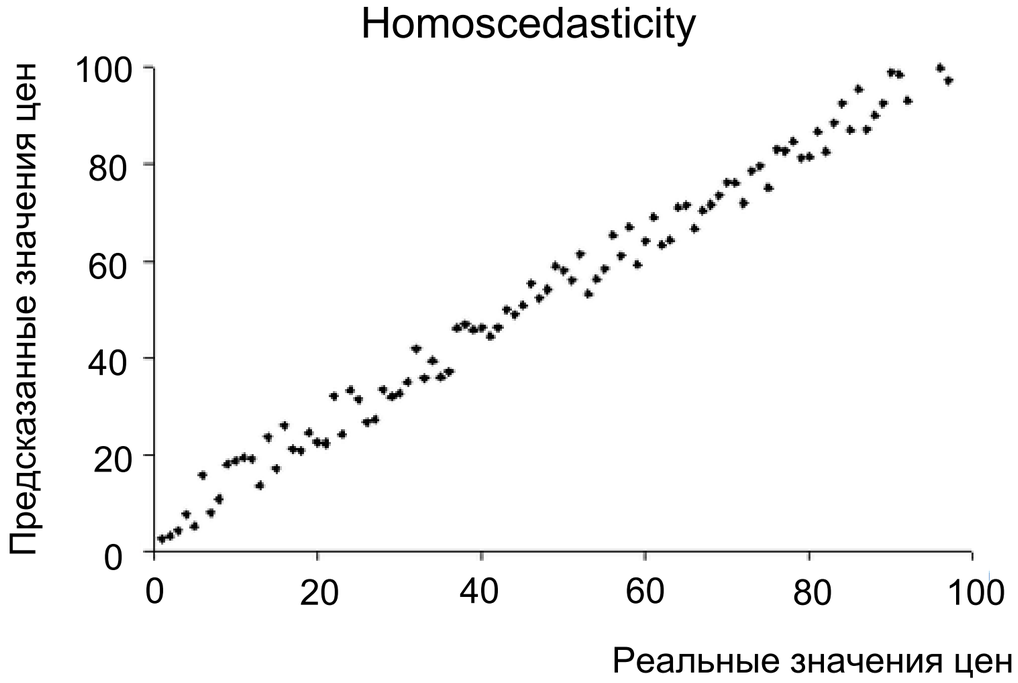
# Гомоскедастичность

Гомоскедастичность — свойство, обозначающее постоянство дисперсии некоторой последовательности случайных величин.

Дисперсия (Variance) — это отклонение точек относительно прямой.

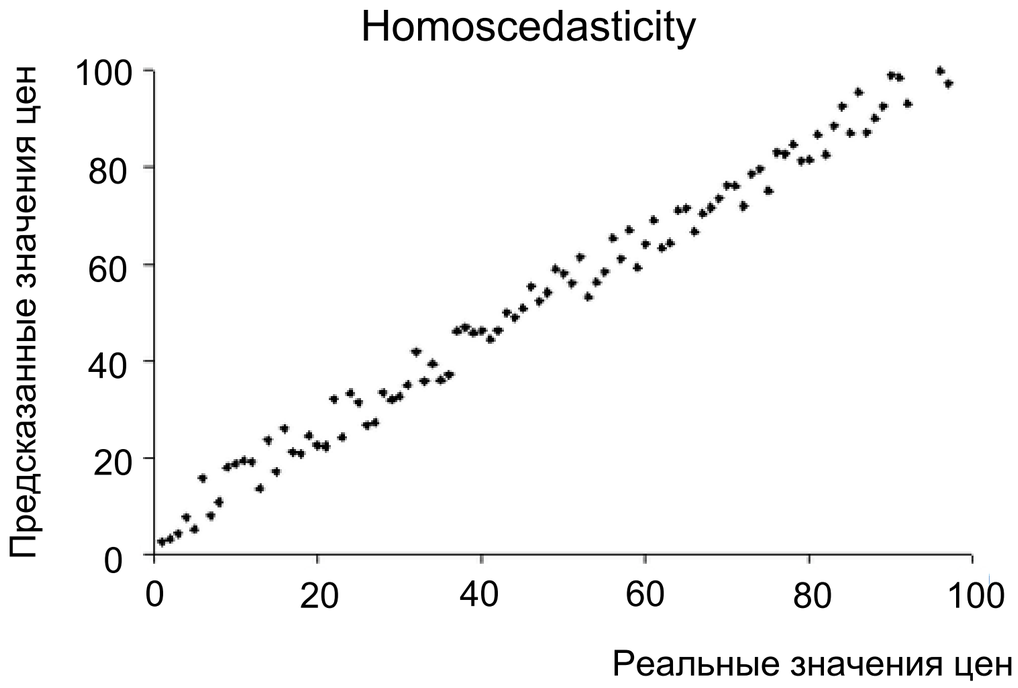
В нашем случае это означает, что дисперсия остатков регрессии должна быть однородной, стабильной для всех наблюдаемых объектов и во все моменты измерения.

# Пример: цена на автомобиль

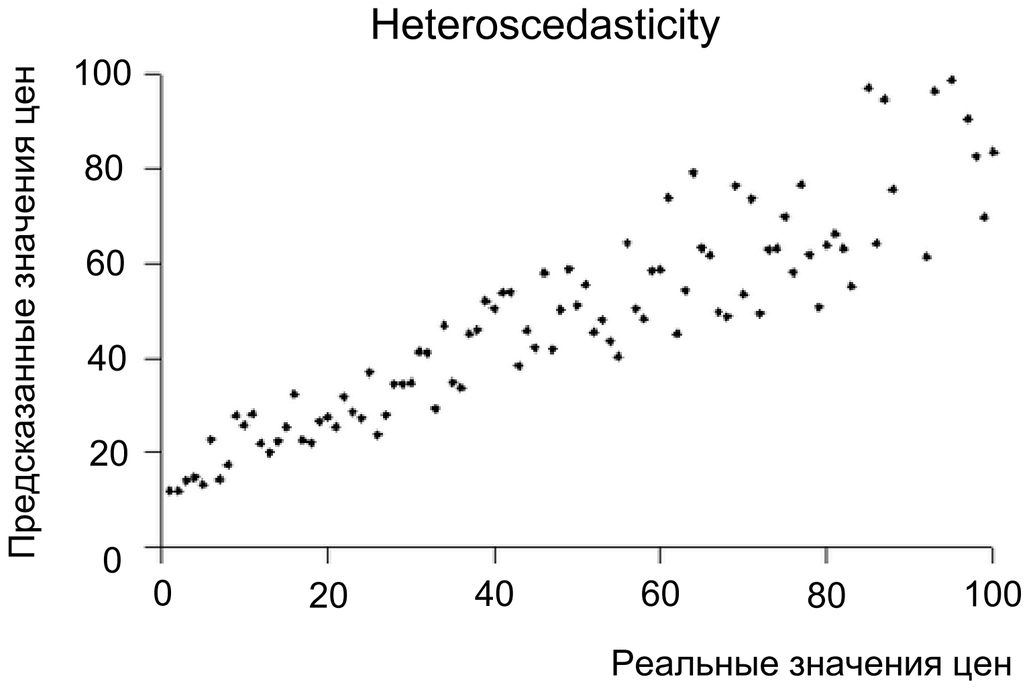



Пример гомоскедастичной вариативности

# Пример: цена на автомобиль




Пример гомоскедастичной вариативности




Пример гетероскедастичности

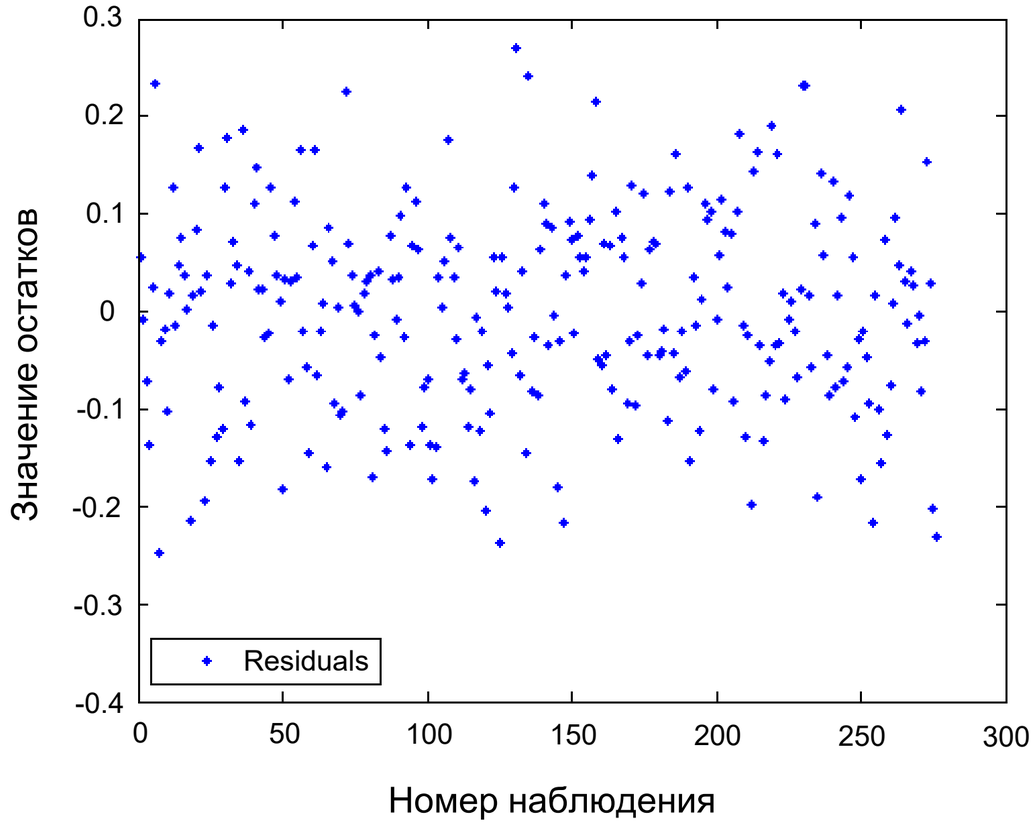
# Гомоскедастичность и гетероскедастичность

Гомоскедастичность — свойство, обозначающее постоянство дисперсии некоторой последовательности случайных величин.

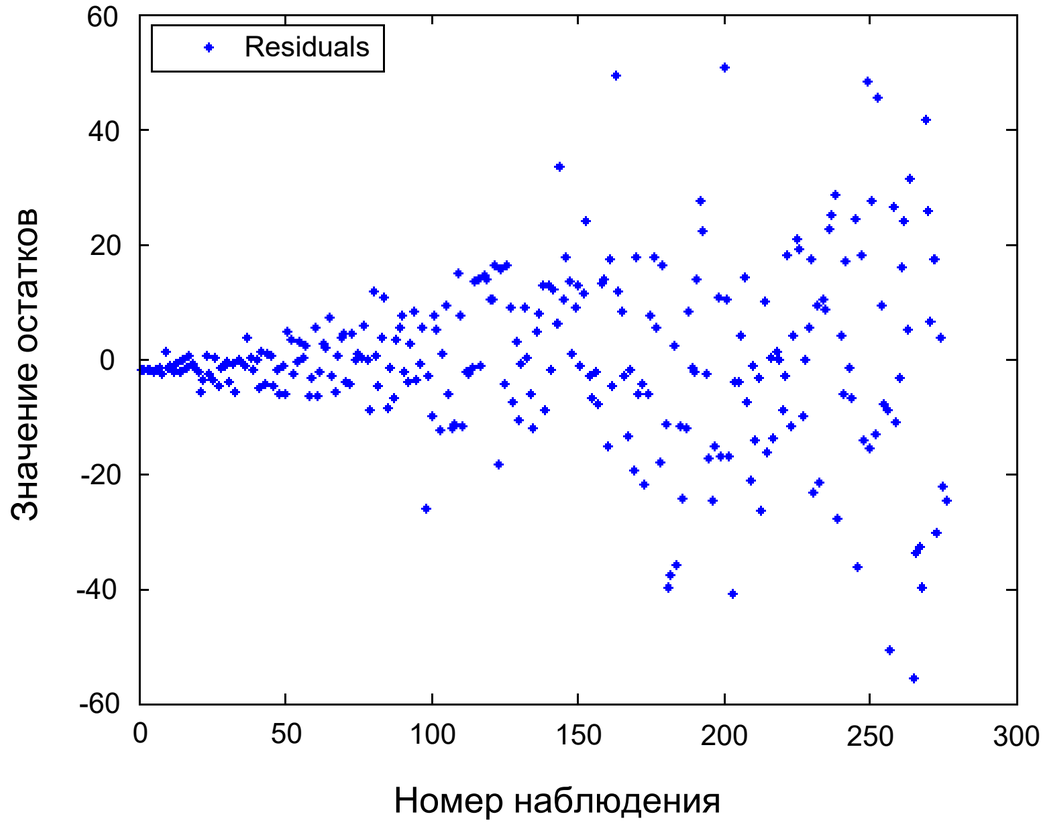
Гетероскедастичность — наоборот, неоднородность наблюдений, выражающуюся в неодинаковой, непостоянной дисперсии случайной ошибки регрессионной модели.

Если выбранная регрессионная модель хорошо описывает истинную зависимость, то остатки должны быть независимыми нормально распределёнными случайными величинами с нулевым средним и в их значениях должен отсутствовать тренд.

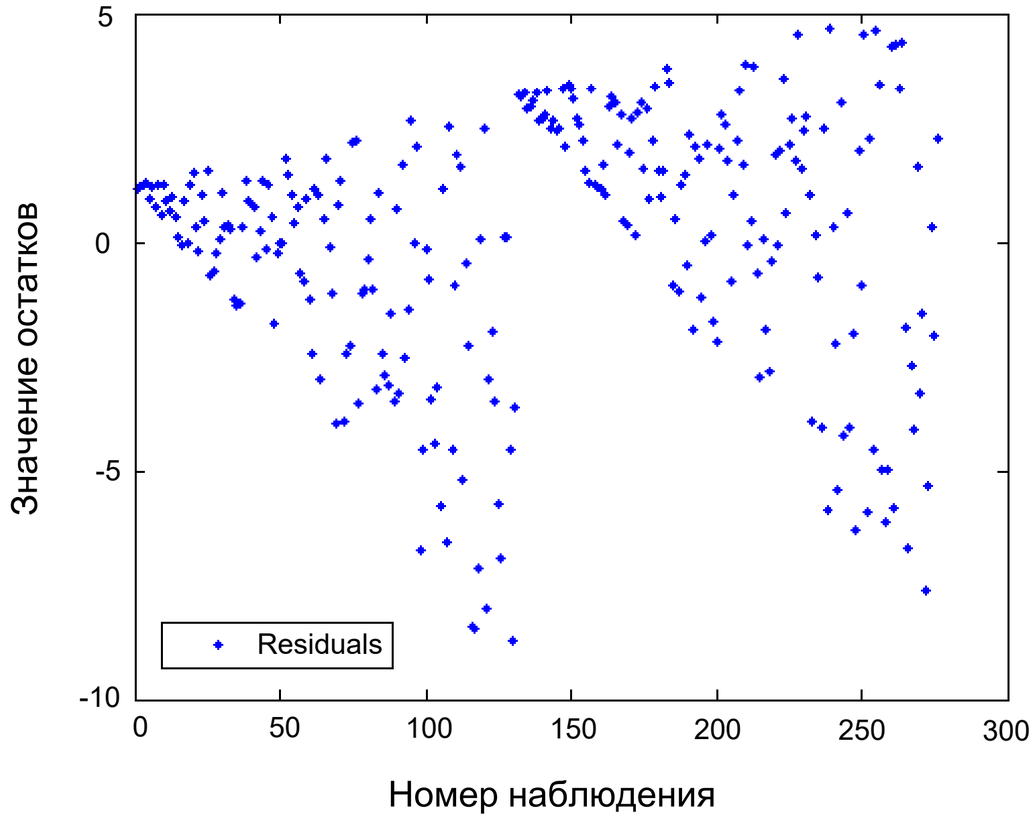
# Гомоскедастичность



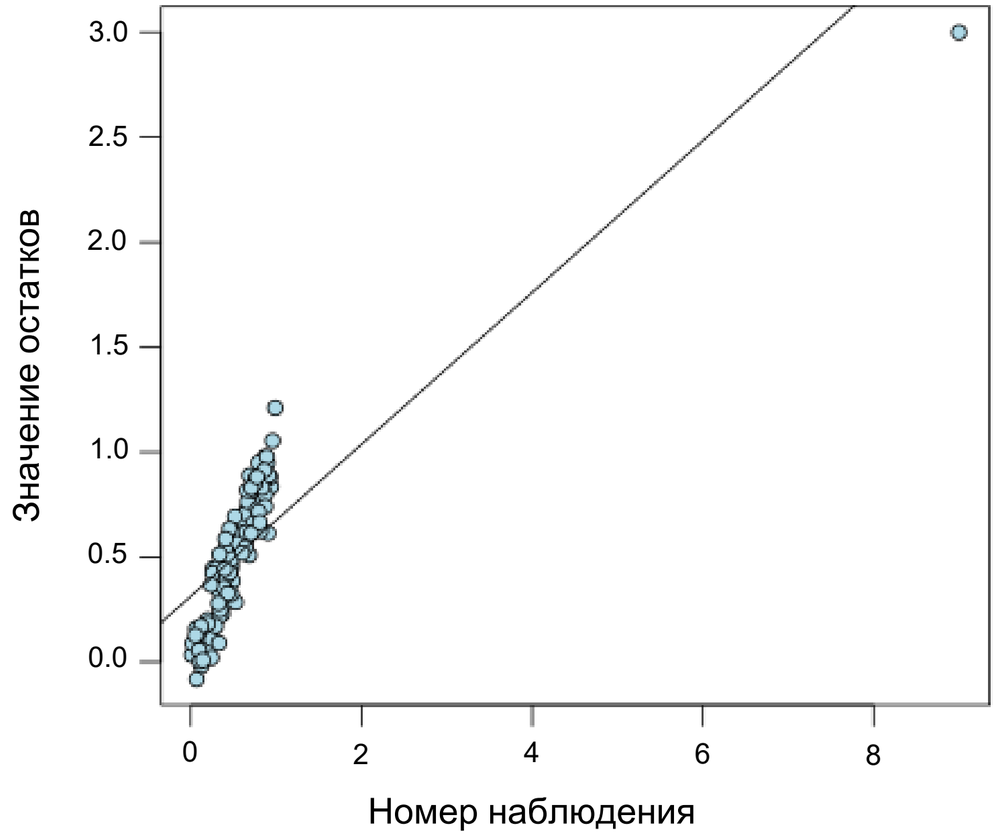
# Гетероскедастичность



# Гетероскедастичность



# Пример с выбросом



# Коэффициент детерминации R2

Коэффициент детерминации принимает значения от 0 до 1.

Чем ближе значение коэффициента к 1, тем сильнее зависимость.

Для приемлемых моделей — 0,5.

Для достаточно хороших моделей — 0,8.

Для идеальной модели — 1.

# Коэффициент детерминации R2

Коэффициент детерминации находится по следующей формуле:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

где $SS_{res} = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$ — уже знакомая сумма квадратов остатков регрессии

а $SS_{tot} = \sum_i (y_i - \hat{y}_{cp})^2$ — общая сумма квадратов,

где $y_{cp} = \frac{1}{n} \sum_i y_i$

# Коэффициент детерминации R2

Расчёт на примере:

пусть

| Y | Y_model |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 3 | 3 |

Сумма квадратов остатков регрессии

$Ss_{res} = (1 - 1)^2 + (2 - 3)^2 + (3 - 3)^2 = 1$

Для вычисления общей суммы квадратов потребуется $y_{cp}$ :

$y_{cp} = (1 + 2 + 3)/3 = 2$

Тогда общая сумма квадратов:

$SS_{tot} = (1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2 = 2$

Коэффициент детерминации равен::

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{1}{2} = 0{,}5$$

# Отбор признаков