

Давайте рассмотрим пример на простых данных, чтобы лучше понять, как работает метод наименьших квадратов (OLS) и остаточная сумма квадратов (RSS).

Пример: Зависимость веса от роста

Предположим, у нас есть данные для трёх людей:

Рост (x) (в см)	Вес (y) (в кг)
150	50
160	60
170	70

Мы хотим найти линейную модель, которая предсказывает вес на основе роста:

$$y = \beta_0 + \beta_1 x$$

1. Расчёт OLS

Метод наименьших квадратов используется для нахождения коэффициентов  $\beta_0$  и  $\beta_1$ , которые минимизируют сумму квадратов ошибок между фактическими значениями и предсказанными значениями.

Матрицы для OLS

Для удобства представим данные в виде матриц:

- Матрица признаков  $X$  (включая столбец с единицами для  $\beta_0$ ):

$$X = \begin{bmatrix} 1 & 150 \\ 1 & 160 \\ 1 & 170 \end{bmatrix}$$

- Вектор целевых значений  $y$ :

$$y = \begin{bmatrix} 50 \\ 60 \\ 70 \end{bmatrix}$$

Формула OLS

Формула для нахождения коэффициентов:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Шаги расчёта:

- Транспонированная матрица  $X^T$ :

$$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 150 & 160 & 170 \end{bmatrix}$$

- Вычисление произведения  $X^T X$ :

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 150 & 160 & 170 \end{bmatrix} \begin{bmatrix} 1 & 150 \\ 1 & 160 \\ 1 & 170 \end{bmatrix} = \begin{bmatrix} 3 & 480 \\ 480 & 77000 \end{bmatrix}$$

- Обратная матрица  $(X^T X)^{-1}$ : Найдём обратную матрицу для  $X^T X$ . Определитель:

$$\det(X^T X) = 3 \cdot 77000 - 480 \cdot 480 = 231000 - 230400 = 600$$

Обратная матрица:

$$(X^T X)^{-1} = \frac{1}{600} \begin{bmatrix} 77000 & -480 \\ -480 & 3 \end{bmatrix} = \begin{bmatrix} 128.33 & -0.8 \\ -0.8 & 0.005 \end{bmatrix}$$

- Вычисление  $X^T y$ :

$$X^T y = \begin{bmatrix} 1 & 1 & 1 \\ 150 & 160 & 170 \end{bmatrix} \begin{bmatrix} 50 \\ 60 \\ 70 \end{bmatrix} = \begin{bmatrix} 50 + 60 + 70 \\ 150 \cdot 50 + 160 \cdot 60 + 170 \cdot 70 \end{bmatrix} = \begin{bmatrix} 180 \\ 25300 \end{bmatrix}$$

5. Вычисление коэффициентов  $\hat{\beta}$ :

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 128.33 & -0.8 \\ -0.8 & 0.005 \end{bmatrix} \begin{bmatrix} 180 \\ 25300 \end{bmatrix} = \begin{bmatrix} -70 \\ 0.7 \end{bmatrix}$$

Итак, получаем:

$$y = -70 + 0.7x$$

## 2. Расчёт RSS

Теперь, когда у нас есть уравнение регрессии, можно рассчитать остаточную сумму квадратов (RSS), чтобы оценить, насколько хорошо модель описывает данные.

RSS рассчитывается по формуле:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Где  $\hat{y}_i$  — предсказанное значение для наблюдения  $i$ :

$$\hat{y}_i = -70 + 0.7x_i$$

Расчёт RSS для каждого наблюдения:

1. Для  $x = 150$ :

$$\hat{y}_1 = -70 + 0.7 \cdot 150 = 35$$

$$(y_1 - \hat{y}_1)^2 = (50 - 35)^2 = 225$$

2. Для  $x = 160$ :

$$\hat{y}_2 = -70 + 0.7 \cdot 160 = 42$$

$$(y_2 - \hat{y}_2)^2 = (60 - 42)^2 = 324$$

3. Для  $x = 170$ :

$$\hat{y}_3 = -70 + 0.7 \cdot 170 = 49$$

$$(y_3 - \hat{y}_3)^2 = (70 - 49)^2 = 441$$

Итоговое значение RSS:

$$RSS = 225 + 324 + 441 = 990$$

## Вывод

1. **OLS** используется для нахождения коэффициентов ( $\beta_0$  и  $\beta_1$ ), которые минимизируют ошибку предсказания.
2. **RSS** — это метрика, показывающая, насколько хорошо модель описывает данные. В нашем случае  $RSS = 990$ .

RSS показывает сумму квадратов разностей между фактическими значениями и предсказанными значениями модели. Чем меньше значение RSS, тем лучше модель описывает данные.

какой вывод сделать можно по этой задаче



По этой задаче можно сделать следующие выводы:

1. **Полученное уравнение регрессии:** В результате метода наименьших квадратов мы получили уравнение линейной регрессии:

$$y = -70 + 0.7x$$

где  $y$  — предсказанный вес, а  $x$  — рост. Коэффициент  $0.7$  показывает, что с увеличением роста на 1 см вес в среднем увеличивается на 0.7 кг. Отрицательный свободный член  $-70$  указывает, что при очень малых значениях роста предсказанный вес будет отрицательным, что говорит о том, что модель не имеет смысла при таких значениях (и требует интерпретации только в разумных пределах).

2. **Качество модели — RSS:** Остаточная сумма квадратов (RSS) составила 990. Это значение показывает, что модель не идеальна и имеет определенные ошибки при описании зависимой переменной. Чем меньше RSS, тем лучше модель описывает данные.
- В нашем случае значение RSS относительно велико (990), что говорит о том, что модель не идеально подходит к данным. Причина может заключаться в том, что модель линейна, а зависимость между весом и ростом может быть сложнее.
3. **Возможные улучшения модели:**
- Многофакторный анализ:** Мы используем только один фактор (рост) для предсказания веса. Возможно, добавление других факторов, таких как возраст, или активность, может улучшить точность модели и уменьшить RSS.
  - Нелинейная регрессия:** В реальной жизни зависимость между ростом и весом может быть нелинейной. Возможно, использование полиномиальной регрессии других нелинейных моделей дало бы более точные результаты.
  - Больше данных:** Мы использовали всего три наблюдения, что является малым количеством данных для построения надежной модели. С увеличением объема данных результаты и качество модели могут улучшиться.
4. **Интерпретация параметров:**
- Коэффициент наклона ( $\beta_1$ ):** Коэффициент  $0.7$  показывает положительную связь между ростом и весом, что логично — с увеличением роста вес также увеличивается.
  - Свободный член ( $\beta_0$ ):** Свободный член  $-70$  может быть трудно интерпретируемым, поскольку предсказывает отрицательные значения веса при очень низких значениях роста. Это говорит о том, что модель не имеет смысла слишком низких значениях роста (например, для детей или младенцев), и нужна осторожность при интерпретации модели за пределами диапазона данных.

### Основной вывод

- Линейная регрессия показывает, что существует прямая зависимость между ростом и весом. Однако модель не идеальна, и ее качество можно улучшить. Использование большего количества данных, добавление дополнительных признаков и применение более сложных моделей могут помочь сделать предсказания более точными и уменьшить ошибку (RSS).
- Значение RSS указывает на наличие значительной ошибки, что может говорить о том, что зависимость между переменными не идеально линейна, и требуется улучшение модели для более точного описания данных.