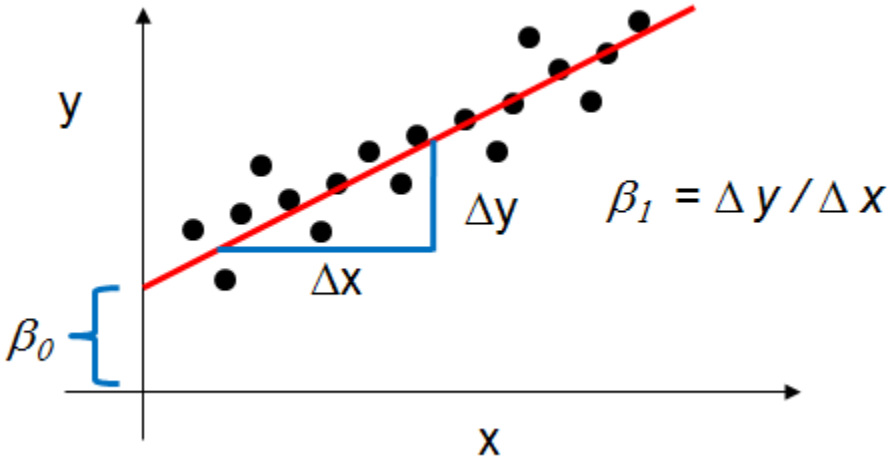
ЛИНЕЙНАЯ РЕГРЕССИЯ

Модель линейной регрессии

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



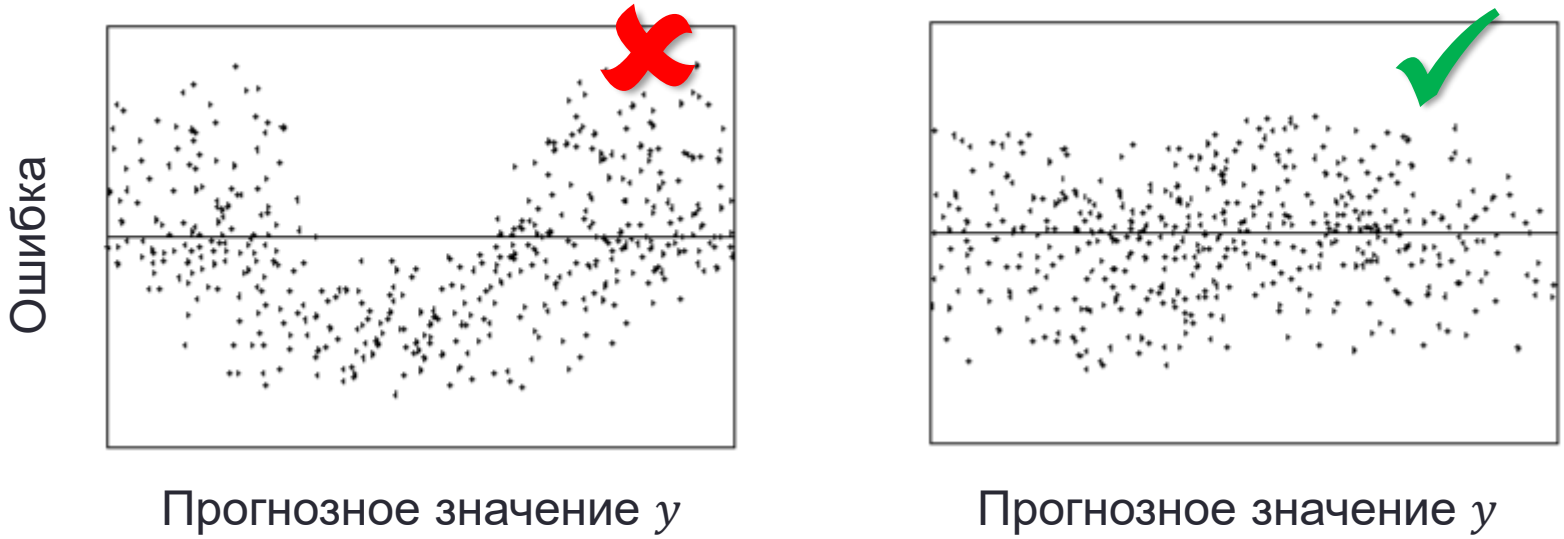
β_0 - константа (англ. intercept)

β_1 - коэффициент (англ. slope, coefficient)

ε – ошибка

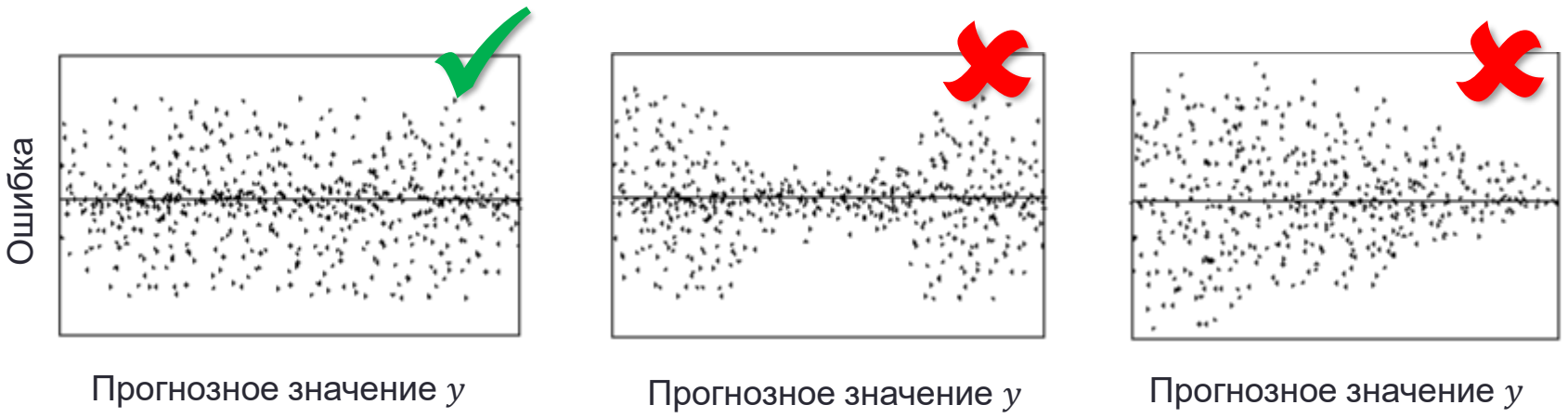
Предположения классической модели линейной регрессии

1. Линейная зависимость зависимых переменных от независимых.



Предположения классической модели линейной регрессии

1. Линейная зависимость зависимых переменных от независимых.
2. Случайные параметры имеют нормальные распределения.
3. Независимые переменные измеряются без ошибок
4. Гомоскедастичность - постоянная или одинаковая дисперсия случайных ошибок модели.



Предположения классической модели линейной регрессии

1. Линейная зависимость зависимых переменных от независимых.
2. Случайные параметры имеют нормальные распределения.
3. Независимые переменные измеряются без ошибок
4. Гомоскедастичность - постоянная или одинаковая дисперсия случайных ошибок модели.
5. Отсутствие автокорреляции ошибок.
6. Отсутствие зависимости между независимыми переменными.

Критерий ошибки в линейной регрессии

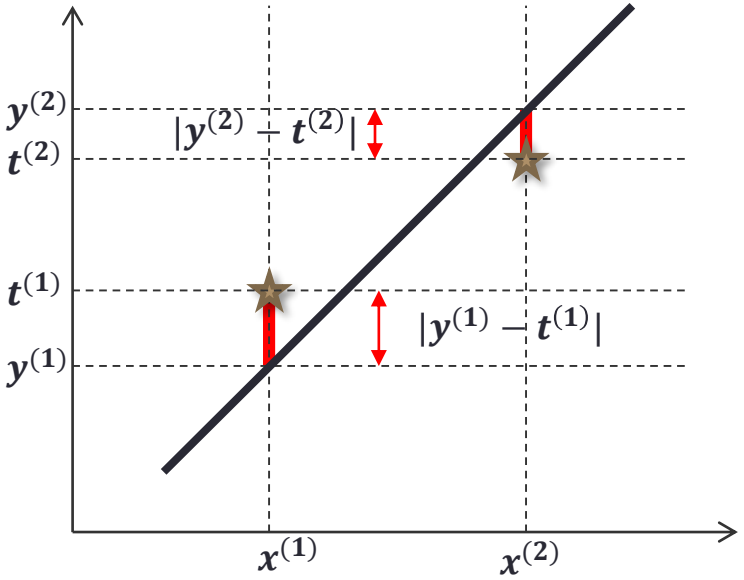
Обучающая выборка

$x^{(1)} \to t^{(1)}$

$x^{(2)} \to t^{(2)}$

…

$x^{(n)} \to t^{(n)}$



Модель

$x^{(1)} \to y^{(1)}$

$x^{(2)} \to y^{(2)}$

…

$x^{(n)} \to y^{(n)}$

Критерий ошибки – сумма квадратов разности фактических и выданных моделью значений:

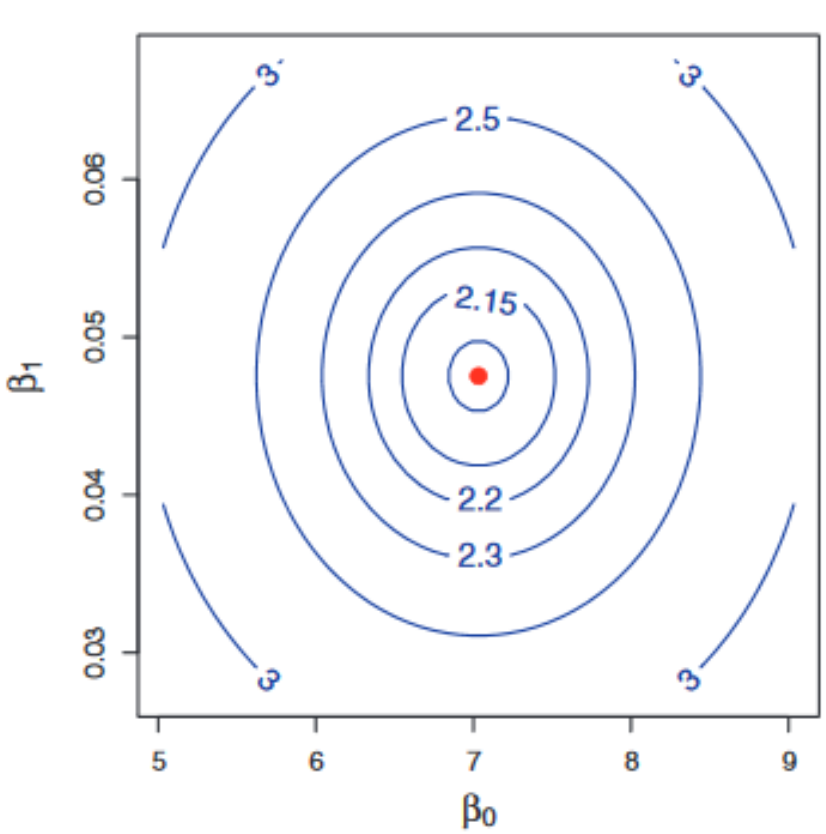
$$RSS = \sum_{i=1}^{n} \left(y^{(i)} - t^{(i)}\right)^2$$

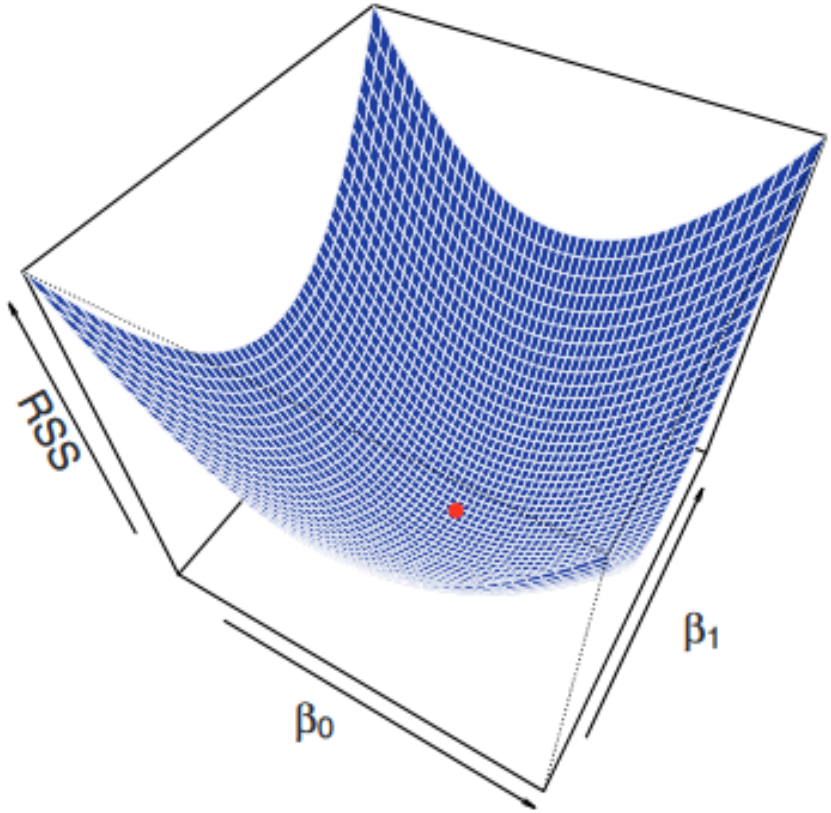
так есть на выходе

так должно быть

Решение задачи линейной регрессии

При выполнении указанных предположений, набор параметров, обеспечивающий минимальное значение ошибки для линейной регрессии может быть найден методом наименьших квадратов (МНК, англ. Ordinary Least Squares, OLS) (Гаусс, 1795, Лежандр 1805).





Оценка качества линейной регрессии

- Ошибка $\boldsymbol{RSS} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(y^{(i)} - \boldsymbol{t}^{(i)}\right)^{\boldsymbol{2}}$
- Часто используют $MSE = \frac{1}{n} RSS$ (Mean Squared Error) и $RMSE = \sqrt{MSE}$ (Root Mean SquaredError)
  - $RMSE \approx$ оценка стандартного отклонения
  - При нормальном распределении случайных значений, большая часть данных лежит не далее двух стандартных отколнений от среднего
  - Соответственно можно считать, что наша точность $\approx \mp 2RMSE$

Коэффициент детерминации

(R-квадрат, coefficient of determination, R-squared)

- Самая простая модель – выход всегда равен своему среднему значению на обучающих данных.
- Наша модель лучше или хуже?

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - y_i)}{\sum_{i=1}^{n}(t_i - \bar{t}\,)}, \qquad \text{где } \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i\,.$$

  - Идеальная модель будет иметь $R^2 = 1$.
  - Модель $y = \bar{t}$ будет иметь $R^2 = 0$.
  - Если $R^2 < 0$ значит модель совсем плохая.

- R^2 это доля дисперсии зависимой переменной, объясняемой нашей моделью.