

Линейная регрессия

Корреляция — статистическая взаимосвязь двух или более случайных величин, при этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{DX \cdot DY}}$$

Коэффициент корреляции:

- может принимать значения от -1 до +1
- знак коэффициента показывает направление связи (прямая или обратная)
- абсолютная величина показывает силу связи

Линейная регрессия

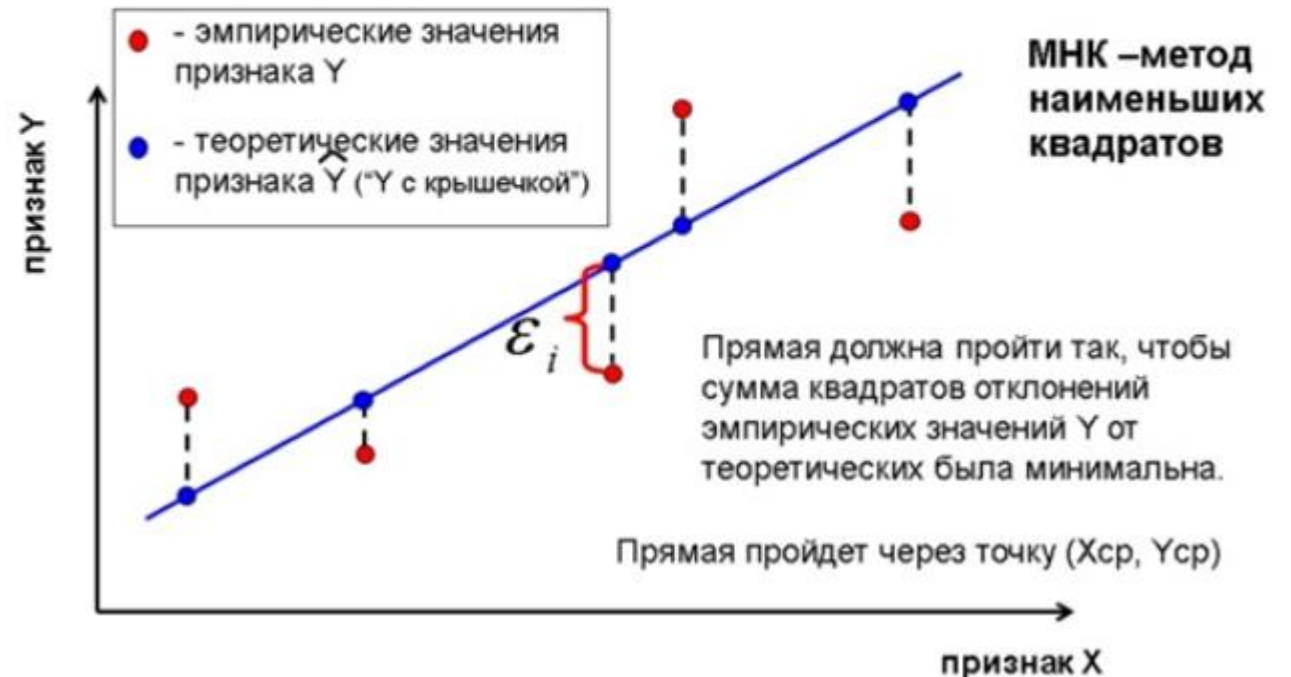
- Строит линейную поверхность в качестве решения
- Ответ – взвешенная сумма значений признаков

$$\hat{y}(w, x) = w_0 + w_1x_1 + \dots + w_px_p$$

Линейная регрессия

Модель – уравнение прямой – $Y = a + b \cdot X$

Построение модели – расчет коэффициентов



Метод наименьших квадратов

- Минимизация $MSE = \sum_{k=0}^n (y_k - \hat{y}_k)^2 / n$
- Условие минимума – равенство нулю производной, откуда
- $w = (X^T X)^{-1} X^T Y$

Линейная регрессия

- Хорошо работает, если зависимость между целевой переменной и признаками имеет линейный характер
- Хорошо интерпретируется
- Перед использованием лучше провести нормализацию данных

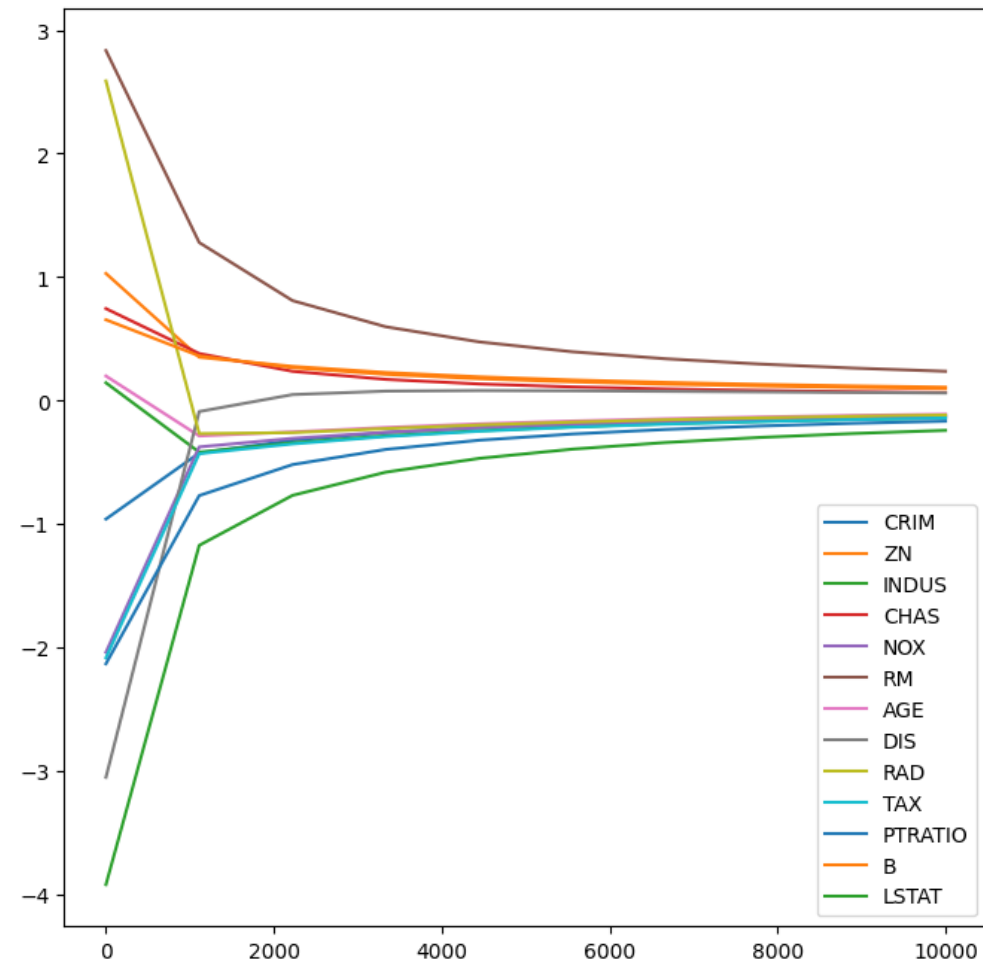
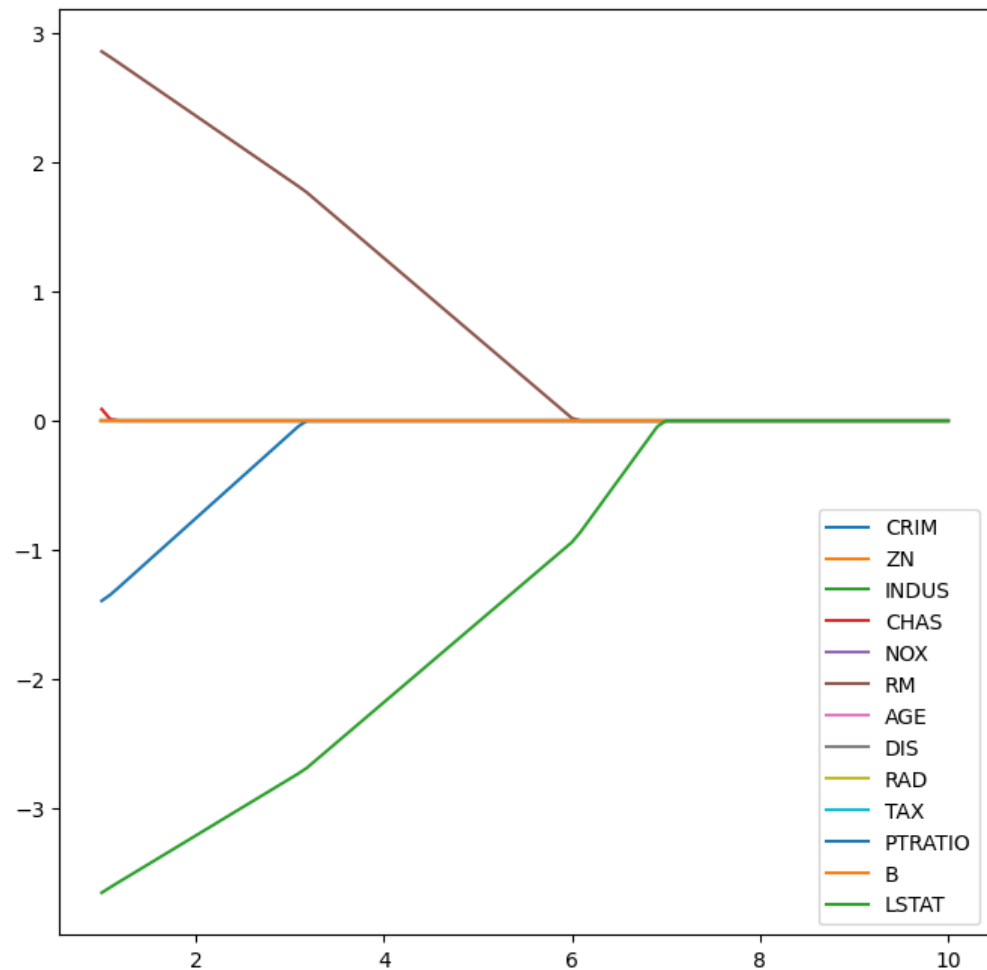
L2 – регуляризация (Ridge-регрессия)

- $L(w)$ – функция потерь
- $L_2 = L(w) + \lambda \sum_{j=0}^s w_j^2$
- λ – параметр регуляризации, коэффициент штрафа модели
- чем крупнее вес, тем он крупнее штрафуются

L1 – регуляризация (Lasso)

- $L(w)$ – функция потерь
- $L_2 = L(w) + \lambda \sum_{j=0}^s |w_j|$

Сравнение L1 и L2



Регуляризация

- Повышает стабильность модели
- Помогает в борьбе с переобучением
- L2 регуляризация имеет более «мягкий» отбор признаков
- Иногда применяется комбинация L1 и L2 – Elastic Net
- Может быть применена к линейной регрессии, логистической регрессии и не только
- Лучше работает с нормализованными признаками