fit, predict, и всё готово

Линейная регрессия

Чертоги разума

Модель:

$$\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad \text{или, что то же самое,} \quad X_i = \sum_{j=1}^{k} z_{ij}\theta_j + \varepsilon_i, \quad i \in \{1, \dots, n\}.$$

где:

- $Z \in \mathbb{R}^{n \times k}$ — *фиксированная* матрица признаков: z_{ij} — j-ый признак для i-ого объекта;
- $\boldsymbol{\theta} \in \mathbb{R}^k$ — вектор неизвестных параметров;
- $\boldsymbol{\varepsilon}$ — случайный n-мерный вектор (шум);
- $\mathbf{X}$ — n-мерный (тоже случайный) вектор целевых величин (таргетов): X_i — таргет i-ого объекта.

Ограничения, накладываемые на модель:

L1 Для всех i выполнено $\mathsf{E}\varepsilon_i = 0$ (в среднем ошибки нет);

L2 Дисперсия ε_i одинакова и равна неизвестному параметру σ^2, причём ε_i попарно нескоррелированы, то есть $\mathsf{D}\boldsymbol{\varepsilon} = \sigma^2 E_n$ (наблюдения друг на друга не влияют).

L3 Столбцы $\mathbf{z}_1, \dots, \mathbf{z}_k$ матрицы Z линейно независимы.

Определение. Оценка $\widehat{\boldsymbol{\theta}} = (Z^T Z)^{-1} Z^T \mathbf{X}$ называется *оценкой по методу наименьших квадратов* (или *МНК-оценкой*).

Свойства МНК-оценки.

- $\mathsf{E}_{\boldsymbol{\theta},\sigma^2}\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, $\mathsf{D}_{\boldsymbol{\theta},\sigma^2}\widehat{\boldsymbol{\theta}} = \sigma^2 (Z^T Z)^{-1}$;
- $\widehat{\boldsymbol{\theta}}$ является лучшей оценкой в среднеквадратичном подходе в классе несмещённых линейных (по $\mathbf{X}$) оценок (*теорема Гаусса-Маркова*);
- Статистика $\widehat{\sigma}^2 = \frac{1}{n-k}\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2$ несмещённо оценивает σ^2.

В модели **гауссовской линейной регрессии**, когда $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, E_n)$, справедлив следующий результат:

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 (Z^T Z)^{-1}), \qquad \frac{1}{\sigma^2}\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2 \sim \chi^2_{n-k}, \quad \widehat{\boldsymbol{\theta}} \perp\!\!\!\perp \widehat{\sigma}^2.$$

На основании этих фактов можно получить статистику Стьюдента для линейной комбинации параметров $\mathbf{c}^T\boldsymbol{\theta}$, по которой построить точный ДИ:

$$\frac{\mathbf{c}^T\widehat{\boldsymbol{\theta}} - \mathbf{c}^T\boldsymbol{\theta}}{\sqrt{\widehat{\sigma}^2 \mathbf{c}^T (Z^T Z)^{-1} \mathbf{c}}} \sim T_{n-k}.$$

Esercizi

1. Пусть в модели линейной регрессии исходный вектор параметров $\boldsymbol{\theta}$ линейно выражается через вектор параметров $\boldsymbol{\psi}$, то есть $\boldsymbol{\theta} = S\boldsymbol{\psi}$, где $S \in GL_k(\mathbb{R})$. Покажите, что $\widehat{\boldsymbol{\theta}} = S\widehat{\boldsymbol{\psi}}$.

2. Пусть $X_1, \ldots, X_n$ — независимые случайные величины, представимые в виде

$$X_i = a + bw_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

где w_i — фиксированные константы. Постройте МНК-оценки для параметров a и b.

3. Недоопределение. В модели линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, где $\mathbf{X} \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n\times k}$, $\theta_i \neq 0$, МНК-оценка $\widehat{\theta}$ вектора $(\theta_1, \ldots, \theta_m)^T$ строится только по первым $m < k$ столбцам матрицы Z (статистик думает, что целевая величина зависит лишь от первых m признаков, когда на самом деле от всех k). Докажите, что полученная оценка может оказаться смещённой.

4. Переопределение. В модели линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, где $\mathbf{X} \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n\times k}$, МНК-оценка $\widehat{\theta}$ вектора $(\theta_1, \ldots, \theta_m)^T$ строится сразу по $m > k$ признакам с матрицей

$$Z' = (\mathbf{z}_1, \ldots, \mathbf{z}_k, \mathbf{z}_{k+1}, \ldots, \mathbf{z}_m) = (Z, \ldots)$$

(статистик думает, что целевая величина зависит от всех m признаков, когда на самом деле только от первых k, то есть $\theta_{k+1} = \ldots = \theta_m = 0$). Докажите, что для всех $i = 1, \ldots, k$ выполнено неравенство $\mathsf{D}_{\boldsymbol{\theta}}\widehat{\theta}_i \geqslant \mathsf{D}_{\boldsymbol{\theta}}\widetilde{\theta}_i$, где $\widetilde{\boldsymbol{\theta}}$ — МНК-оценка, построенная по матрице Z (иными словами, добавление бесполезных признаков увеличило дисперсию оценок действительно нужных параметров).

5. Пусть $X_1, \ldots, X_n \sim \mathcal{N}(a, \sigma^2)$ и $Y_1, \ldots, Y_m \sim \mathcal{N}(b, \sigma^2)$ — независимые выборки. Сведя задачу к линейной регрессии, постройте точный доверительный интервал для $a - b$.

6. Постройте точный доверительный интервал уровня доверия γ для целевой величины нового объекта с признаками $\mathbf{z}_0$ (то есть $\mathbf{z}_0^T\boldsymbol{\theta} + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ не зависит от $\boldsymbol{\varepsilon}$) для модели гауссовской линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$.

Замечание. ДИ для случайной величины называют *предсказательным интервалом.*

7. Докажите, что статистика $(\widehat{\boldsymbol{\theta}}, \|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2)$ является **(а)** достаточной; **(б)** полной в модели гауссовской линейной регрессии.

8. В модели линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, где $\mathsf{E}\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathsf{D}\boldsymbol{\varepsilon} = \sigma^2 V$ (V — п.о. матрица) найдите лучшую оценку среди линейных несмещённых в среднеквадратичном подходе.