

Семинар - 9.

Линейная регрессия. Доверительные интервалы

Гауссовская линейная регрессия.

$$y = \sum_{j=1}^p \psi_j(x) \theta_j + \varepsilon_i = \boldsymbol{\psi}(x)^T \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Пусть имеются данные $\{(X_i, Y_i) : 1 \leq i \leq n\}$, причем наблюдения $Y_1, \dots, Y_n \in \mathbb{R}$ получены из модели

$$\mathbf{Y} = \Psi^T \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n),$$

где $\boldsymbol{\theta} \in \mathbb{R}^p$, $\Psi = (\boldsymbol{\psi}(X_1), \dots, \boldsymbol{\psi}(X_n)) \in \mathbb{R}^{p \times n}$ — матрица плана полного ранга, то есть $\text{rank}(\Psi) = p$, $X_1, \dots, X_n$ детерминированные, а $\varepsilon_1, \dots, \varepsilon_n$ — независимые центрированные случайные величины с дисперсией σ^2 . Мы предполагаем, что матрица $\Psi \in \mathbb{R}^{p \times n}$ имеет полный ранг,

Оценка в методе наименьших квадратов (ОМНК):

$$\hat{\boldsymbol{\theta}}_{LSE} = \arg \min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n (Y_i - \boldsymbol{\psi}(X_i)^T \boldsymbol{\theta})^2 = \arg \min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{Y} - \Psi^T \boldsymbol{\theta}\|^2 = (\Psi \Psi^T)^{-1} \Psi \mathbf{Y}.$$

Введем ортогопроектор на линейное подпространство размерности p , натянутое на строки матрицы Ψ : $(\boldsymbol{\psi}_j(X_1), \dots, \boldsymbol{\psi}_j(X_n))^T \in \mathbb{R}^n$, $1 \leq j \leq p$

$$\Pi = \Psi^T (\Psi \Psi^T)^{-1} \Psi \in \mathbb{R}^{n \times n}.$$

Тогда оценка регрессионной функции $\boldsymbol{\psi}(x)^T \boldsymbol{\theta}$ в точках $X_1, \dots, X_n$:

$$\hat{f} = \Psi^T \hat{\boldsymbol{\theta}}_{LSE} \equiv \Pi \mathbf{Y}.$$

Утверждение 1 (Геометрическая интерпретация $\hat{f}$). $\hat{f}$ является проекцией вектора наблюдений на линейную оболочку строк матрицы плана Ψ , то есть на $\text{Im}(\Psi^T)$.

Лемма (Свойство гауссовских векторов). Пусть вектор X имеет нормальное распределение $\mathcal{N}(\mu, \Sigma_n)$. Тогда для произвольной матрицы $B \in \mathbb{R}^{m \times n}$ вектор BX имеет нормальное распределение $\mathcal{N}(B\mu, B\Sigma_n B^T)$.

Лемма. Пусть вектор X имеет стандартное нормальное распределение $\mathcal{N}(0, I_n)$ и пусть B — ортогопроектор. Тогда $Q = X^T B X$ имеет распределение хи-квадрат $\chi^2(\text{tr } B)$.

Задачи

1. Найти оценку неизвестного параметра $\boldsymbol{\theta}$ методом максимального правдоподобия. Будет ли она несмещенной? Будет ли она оптимальна? Сравнить с оценкой наименьших квадратов в модели линейной регрессии (то есть когда вектор ошибок не обязательно гауссовский).
2. Найти оценку максимального правдоподобия для σ^2 в гауссовской линейной регрессии. Будет ли она несмещенной?
3. Доказать, что $\frac{1}{\sigma^2} \|\Pi \boldsymbol{\varepsilon}\|^2$ имеет распределение $\chi^2(p)$. — не ученик
4. Доказать, что $\hat{\boldsymbol{\theta}}_{LSE}$ и $\mathbf{Y} - \Psi^T \hat{\boldsymbol{\theta}}_{LSE}$ независимы. Доказать, что $\frac{1}{\sigma^2} \|\mathbf{Y} - \Psi^T \hat{\boldsymbol{\theta}}\|^2$ имеет распределение хи-квадрат $\chi^2(n-p)$. — не ученик.

5. Постройте точный $1 - \alpha$ -доверительный интервал для θ_j , где $j = \{1, \dots, p\}$, если
 - (a) σ^2 известен;
 - (b) σ^2 неизвестна.
6. Постройте точное $1 - \alpha$ -доверительное множество для $\boldsymbol{\theta}$, если σ^2 неизвестна.
7. Постройте точный $1 - \alpha$ -доверительный интервал для σ^2 .

1. Найти оценку неизвестного параметра θ методом максимального правдоподобия. Будет ли она несмещенной? Будет ли она оптимальна? Сравнить с оценкой наименьших квадратов в модели линейной регрессии (то есть когда вектор ошибок не обязательно гауссовский).

$$y = \psi^T \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

↑ X результаты

$$y_1 = \psi(X_1)^T \theta + \varepsilon_1$$

↑ ↑
 $\mathbb{R}^p$ $\mathbb{R}^p$

$$= \sum_{j=1}^p \psi_j \theta_j$$

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_p \end{pmatrix} = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^p \end{pmatrix}$$

$$\psi \in \mathbb{R}^{p \times n}$$

$$= \mathbb{R}^{p \times L}$$

$$X_1, \dots, X_n \sim P_\theta$$

$$(y_1, \dots, y_n) \sim \mathcal{N}(\psi^T \theta, \sigma^2)$$

$$L(\theta) = \prod_{i=1}^n p_\theta(y_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \psi(X_i)^T \theta)^2}{2\sigma^2}}$$

$$y_i \sim \psi(X_i)^T \theta + \mathcal{N}(0, \sigma^2) \sim \mathcal{N}(\underbrace{\psi(X_i)^T \theta}_\mu, \sigma^2)$$

$$\ln L(\theta) = \sum_{i=1}^n \left(\underbrace{\ln \frac{1}{\sqrt{2\pi}\sigma}}_C - \frac{(y_i - \psi(X_i)^T \theta)^2}{2\sigma^2} \right)$$

$$\hat{\theta} = ?$$

$$= nC - \frac{1}{2\sigma^2} \sum (y_i - \psi(X_i)^T \theta)^2$$

$$= nC - \frac{1}{2\sigma^2} \|y - \psi^T \theta\|_2^2 \quad \leftarrow \text{перешли в векторной форме}$$

$$df = \langle \nabla f(\theta), d\theta \rangle$$

$$L(\theta) = \text{const} \cdot \frac{1}{\sigma^2} \langle y - \psi^T \theta, y - \psi^T \theta \rangle$$

$$dL(\theta) = \frac{1}{\sigma^2} \langle \psi^T d\theta, y - \psi^T \theta \rangle$$

$$= \langle \underbrace{\psi(y - \psi^T \theta)}_{\sigma^2 \nabla f(\theta)}, d\theta \rangle$$

$$\langle Ax, y \rangle = \langle x, A^T y \rangle$$

$$\left. \frac{\partial f(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} \stackrel{\theta\text{-максим.}}{=} \Psi(Y - \Psi^T \hat{\theta}) = 0$$

$$\Psi Y = (\Psi \Psi^T) \hat{\theta} \Rightarrow (\Psi \Psi^T)^{-1} \Psi Y = \hat{\theta}$$

Несмещенность:

$$(\Psi \Psi^T)^{-1} \Psi E Y = \theta$$

2. Найти оценку максимального правдоподобия для σ^2 в гауссовской линейной регрессии. Будет ли она несмещенной?

$$\ln L(\theta) = \sum_{i=1}^n \left(\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(Y_i - \Psi(X_i)^T \theta)^2}{2\sigma^2} \right) = \frac{n}{2} \ln \left(\frac{1}{2\pi\sigma^2} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \Psi(X_i)^T \theta)^2$$

$$\frac{\partial}{\partial \sigma^2} \ln L(\theta) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \|Y - \Psi^T \theta\|_2^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n} \|Y - \Psi^T \theta\|_2^2$$

незав. от θ , посчит. оценку.

$$\hat{\sigma}^2 = \frac{1}{n} \|Y - \Psi^T \theta\|_2^2 = \frac{1}{n} \|Y - \Pi Y\|_2^2 = \frac{1}{n} \|(I - \Pi)Y\|_2^2$$

$$\Pi = \Psi (\Psi^T \Psi)^{-1} \Psi^T$$

5. Постройте точный $1 - \alpha$ -доверительный интервал для θ_j , где $j = \{1, \dots, p\}$, если

- (a) σ^2 известен;
- (b) σ^2 неизвестна.

$\hat{\theta} \sim \mathcal{N}((\Psi \Psi^T)^{-1} \Psi^T \theta, (\Psi \Psi^T)^{-1} \sigma^2)$
 $y \sim \mathcal{N}(\Psi^T \theta, \sigma^2 I_n)$

$G = (\hat{\theta}_j - \theta_j) \frac{1}{\sqrt{(\Psi \Psi^T)^{-1}_{jj} \sigma^2}} \sim \mathcal{N}(0, 1)$

$\hat{\theta}_j \sim \mathcal{N}(\theta_j, (\Psi \Psi^T)^{-1}_{jj} \sigma^2)$

б) σ^2 — не известна.

$\hat{\sigma}^2 = \frac{1}{n-p} \|(I - \Pi)y\|_2^2$

$\frac{\hat{\theta}_j - \theta_j}{\sqrt{(\Psi \Psi^T)^{-1}_{jj} \hat{\sigma}^2}} \sim \mathcal{N}(0, 1)$

$\frac{1}{\hat{\sigma}^2} \|(I - \Pi)y\|_2^2 \sim \chi^2(n-p)$

$\text{trace} = 1$ $\text{trace } p$

задача

см. теор. с. 11

убо. расч.

Стьюдент.