Св-ва оценок в модели лин. регр. при несмещ. и гомоскедастичности шума

3.3

Предпологаемая зависимость

$$y(x) = \theta^T X$$

Наблюдаемая зависимость

$$y(x) = X\theta + \varepsilon$$

$y \in \mathbb{R}^n$ — случ.

$X \in \mathbb{R}^{n \times d}$ — известен, неслуч.

$\theta \in \mathbb{R}^d$ — неизв.

$\varepsilon \in \mathbb{R}^n$ — случ. неизв.

Задача: оценить Θ

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - \theta^T x_i)^2 = \|y - X\theta\|^2$$

residal sum of squares

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \; RSS(\theta) \quad \text{— МНК оценка}$$

Если $X^T X$ — не вырр., то $\hat{\theta} = (X^T X)^{-1} X^T Y$

Свойства:

1) Если $E\varepsilon = 0$, то $E\hat{\theta} = \theta$

$$E\hat{y}(x) = y(x), \text{ где } \hat{y}(x) = \hat{\theta}^T X$$

[несмещённость шума]

2) Если $D\varepsilon = \sigma^2 I_n$, $E\varepsilon = 0$, то

$$D\hat{\theta} = \sigma^2 (X^T X)^{-1}$$

$$D\hat{y}(x) = \sigma^2 x^T (X^{-1} X)^{-1} x$$

[гомоскедастичность шума]

$\det(X^{-1}X) \sim 0 \implies$ большая дисперсия $\Rightarrow$ больно

$$\boxed{\underset{n\times d}{D\,A}\,\underset{d\times 1}{\xi} = \underset{n\times d}{A}\,\underset{d\times d}{D\xi}\,\underset{d\times n}{A^T}}$$

▲ $$E\hat{\theta} = E\left((X^{-1}X)^{-1}X^T Y\right) = (X^T X)^{-1} X^T E(X\theta + \varepsilon) =$$

$$= (X^T X)^{-1} X^T X \theta = \theta$$

$$D\hat{\theta} = D\left((X^{-1}X)^{-1}X^T Y\right) = (X^T X)^{-1} X^T \, D(X\theta + \varepsilon) \, X(X^T X)^{-1} =$$

$$= (X^T X)^{-1} \sigma^2 I_n X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

Утв. Если $E\varepsilon = 0$ $D\varepsilon = \sigma^2 I_n$, то

оценка $\hat{\sigma}^2 = \dfrac{RSS(\hat{\theta})}{n-d} = \dfrac{\| y - X\hat{\theta}\|^2}{n-d}$ — несмещ. оценка σ^2

▲ $E\,RSS(\hat{\theta}) = E\sum\limits_{i=1}^{n} (y_i - \hat{\theta}^T x_i)^2 = D\sum\limits_{i=1}^{n} (y_i - \hat{\theta}^T x_i) =$

$$\left[\begin{array}{l} E\,(y_i - \hat{\theta}^T x_i) = 0 \\ \quad\ \ \| \\ \quad\ \theta^T x_i + \varepsilon \end{array} \right]$$

$= \operatorname{tr} D(y - X\hat{\theta})$

$$D(y - X\hat{\theta}) = D\Big[\big(I_n - \underbrace{X(X^TX)^{-1}X^T}_{A\text{ — симметр.}}\big)y\Big] = D\big[(I_n - A)y\big] =$$

$\left[\begin{array}{l} A \text{ обратима} \Rightarrow \varepsilon \equiv 0 \\ n < d \Rightarrow rk\, A < n \Rightarrow A \text{ необратима} \end{array}\right.$

$$= (I_n - A)\, D y\, (I_n - A)^T = (I_n - A)\, \sigma^2 I_n\, (I_n - A)^T =$$

$$= \sigma^2 (I_n - A^T - A + AA^T) =$$

$$\left[AA^T = X\cancel{(X^TX)^{-1}X^TX}(X^TX)^{-1}X^T = A\right]$$

$$= \sigma^2 I_n (I_n - A)$$

$$\left[\operatorname{tr} A = \operatorname{tr}\left(X(X^TX)^{-1}X^T\right) \underset{\text{св-во следа}}{=} \operatorname{tr}\left(X^TX(X^TX)^{-1}\right) = \operatorname{tr}(I_d) = d\right]$$

$$\Rightarrow \operatorname{tr} D(y - X\hat{\theta}) = \operatorname{tr}\left(\sigma^2(I_n - A)\right) = \sigma^2(n - d)$$

Пример $X_1 \ldots X_n \sim Exp(\theta)$ [когда-то было]

$\hat{\theta}_1 = 1/\overline{X}$

$\hat{\theta}_2 = -\ln \overline{I\{X > 1\}}$

сильносост. и ас. норм. оценки θ

Какая лучше?

Сравнение оценок

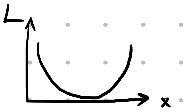
3.4

Хотим оценить $\tau(\theta) \in \mathbb{R}^d$

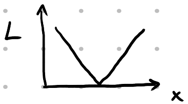
Опр. Ф-ция $L: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, характеризующая степень отклонения оценки от $\tau(\theta)$ наз-ся *функцией потерь*

Примеры:

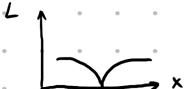
1) $d=1$ $\quad L(x,y) = (x-y)^2$ — квадратичная



2) $d=1$ $\quad L(x,y) = |x-y|$ — абсолютная



3) $d=1$ $\quad L(x,y) = \ln(1+|x-y|)$



4) $d > 1$ $\quad L(X,Y) = (x-y)^T A (x-y)$,

где A – сим. неотр. полуопр.

$A = I_d$, тогда $h(x,y) = \sum_{i=1}^{d} (x_i - y_i)^2$

Если $\hat{\theta}$ — оценка $\tau(\theta)$, то при таком оценивании

$$L(\hat{\theta}, \tau(\theta)) \text{ — штраф}$$

При таком подходе есть недостаток:

штраф случаен, при разных реализациях получаем разные штрафы

Опр. Ф-ция риска оценки $\hat{\theta}$ величины $\tau(\theta)$

$$R_{\hat{\theta}, \tau(\theta)}(\theta) = E_\theta L(\hat{\theta}, \tau(\theta)) \quad \leftarrow \text{ф-ция от } \theta$$

(risk function)

Если $L(x, y) = (x-y)^2$ - кв. ф-ция потерь, то

$$MSE_{\hat{\theta}, \tau(\theta)} = E_\theta (\hat{\theta} - \tau(\theta))^2$$ - средняя квадратичная ошибка

$$MAE_{\hat{\theta}, \tau(\theta)} = E_\theta |\hat{\theta} - \tau(\theta)|$$ - средняя абсолютная ошибка

Замечание:

Если $\tau(\theta) = \theta$, то τ будем опускать

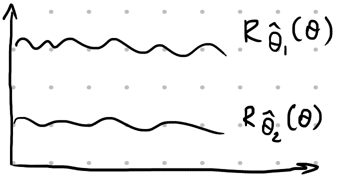
Пример $X_1 \dots X_n$ — выборка $EX_1 = \theta$

$DX_1 < +\infty$

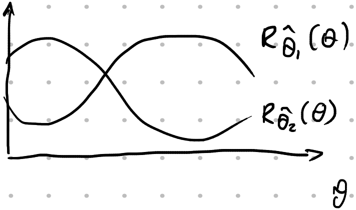
Посчитаем MSE для $\underset{\hat{\theta}_1}{\overset{\parallel}{X_1}}$, $\underset{\hat{\theta}_2}{\overset{\parallel}{\overline{X}}}$ — оц. θ

$$MSE_{\hat{\theta}_1}(\theta) = E_\theta(X_1 - \underbrace{\theta}_{EX_1})^2 = DX_1$$

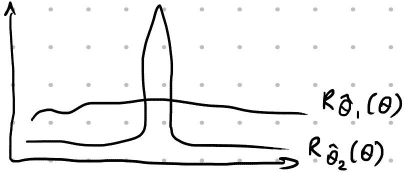
$$MSE_{\hat{\theta}_2}(\theta) = E_\theta(\overline{X} - \underbrace{\theta}_{E\overline{X}})^2 = D\overline{X} = \frac{DX_1}{n}$$



$R_{\hat{\theta}_1}(\theta)$
$R_{\hat{\theta}_2}(\theta)$



$R_{\hat{\theta}_1}(\theta)$
$R_{\hat{\theta}_2}(\theta)$
ϑ



$R_{\hat{\theta}_1}(\theta)$
$R_{\hat{\theta}_2}(\theta)$

Что лучше?

Подходы к сравнению оценок

1) Равномерный подход

- $\hat{\theta}_1$ не хуже $\hat{\theta}_2$, если $\forall \theta \in \Theta$ $R_{\hat{\theta}_1}(\theta) \leq R_{\hat{\theta}_2}(\theta)$
- $\hat{\theta}_1$ лучше $\hat{\theta}_2$, если $\hat{\theta}_1$ не хуже $\hat{\theta}_2$

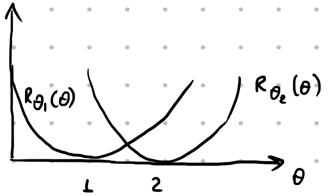
  и $\exists \theta \in \Theta$ $R_{\theta_1}(\theta) < R_{\theta_2}(\theta)$
- $\hat{\theta} \in \mathcal{K}$ – класс оценок, $\hat{\theta}$ – наилучшая в $\mathcal{K}$, если

  она лучше любой другой
- Если $L(x, y) = (x-y)^2$ – ф-ция потерь

  то говорят о среднеквадратичном подходе

Утв. В классе $\mathcal{K}$ может не быть наилучшей оценки.

Пример $\Theta = \mathbb{R}$ $\qquad \mathcal{K} = \{\hat{\theta}_1 = 1, \ \hat{\theta}_2 = 2\}$



Утв. MSE допускает bias-variance разложение

$$MSE_{\hat{\theta},\tau}(\theta) = E_\theta(\hat{\theta} - \tau(\theta))^2 = \underbrace{D\hat{\theta}}_{\text{variance}} + \underbrace{(E_\theta\hat{\theta} - \tau(\theta))^2}_{\text{bias}^2}$$

Следствие:

наилучший по MSE в классе $\mathcal{K} = \{$ несмещ. оц. $\}$ будет оценка с наим дисперсией

$\blacktriangle$

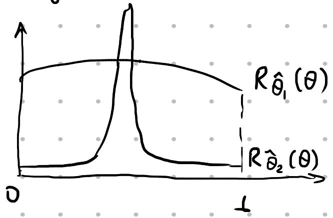
$$E_\theta(\hat{\theta} - \tau(\theta))^2 = \underbrace{E_\theta(\hat{\theta} - E_\theta\hat{\theta})^2}_{D\hat{\theta}} + E(\underbrace{E_\theta\hat{\theta} - \tau(\theta)}_{\text{const}})^2 + 2\overbrace{E_\theta(\hat{\theta} - E_\theta\hat{\theta})}^{0 \,=\, \text{т.к. несмещ.}}\overbrace{(\hat{\theta} - \tau(\theta))}^{\text{const}} =$$

$$= D\hat{\theta} + (E_\theta\hat{\theta} - \tau(\theta))^2$$

2) Байесовский подход

Пусть Q – распределение на Θ

- $\hat\theta_1$ не хуже $\hat\theta_2$, если $E_Q R_{\hat\theta_1, \tau}(\theta) \leq E_Q R_{\hat\theta_2, \tau}(\theta)$



$Q = U[0, 1]$

при таком Q мы по факту считаем интеграл

$\hat\theta_2$ лучше $\hat\theta_1$

3) Минимаксный подход

- $\hat{\theta}_1$ не хуже $\hat{\theta}_2$, если $\sup\limits_{\theta \in \Theta} R_{\hat{\theta}_1}(\theta) \leq \sup\limits_{\theta \in \Theta} R_{\hat{\theta}_2}(\theta)$

4) Асимптотический подход (для ас. норм. оц., $\mathbb{R}$)

$\hat{\theta}_1, \hat{\theta}_2$ — а.н.о. с ас.д. σ_1^2, σ_2^2

- $\hat{\theta}_1$ не хуже $\hat{\theta}_2$, если $\forall \theta \in \Theta \quad \sigma_1^2(\theta) \leq \sigma_2^2(\theta)$
- $\hat{\theta}_1$ лучше $\hat{\theta}_2$, если $\downarrow + \exists \theta \in \Theta \quad \sigma_1^2(\theta) < \sigma_2^2(\theta)$
- $ARE^{\tau}_{\hat{\theta}_1, \hat{\theta}_2} = \dfrac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}$ относительная асимптотическая эффективности

показывает насколько $\hat{\theta}_1$ лучше $\hat{\theta}_2$

Опр $\hat{\theta}$ наз-ся асимптотически эффективной, если она наилучшая в классе всех а.н.о. с непр. а.д. в асимпт. подходе

вроде без непр не будет наил. оц. в классе

[т.е. наим. ас. дисп. среди всех а.н.о.]

[L1 — L8] ОМП — ас. эф. оценка, т.е. $i(\theta)^{-1}$ — наим ас. д.

матрички сравнивать страшно

этого мы делать не будем

Пример $X_1 - X_n \sim N(\theta, 1)$

м/б здесь $\frac{\pi^2}{4}$

ЦПТ: $\overline{X}$ — а.н.о. θ с а.д. $\sigma^2(\theta)$

т. о выборочной медиане $\hat{\mu}$ — а.н.о. θ с ас.д. $\sigma_2^2(\theta) = \frac{\pi}{2}$

$ARE_{\overline{X}, \mu} = \frac{\pi}{2} \approx 1.57$, т.е. $\overline{X}$ в ~ 1.57 раза лучше μ

Пример $X_1 \ldots X_n \sim Exp(\theta)$

[раньше было] $\hat{\theta}_1 = \frac{1}{\overline{X}}$ а.н.о. с а.д. $\sigma_1^2(\theta) = \theta^2$

$\hat{\theta}_2 = -\ln \overline{I\{X > 1\}}$ а.н.о. с а.д. $\sigma_2^2(\theta) = e^\theta - 1$

$$ARE_{\hat{\theta}_1, \hat{\theta}_2} = \frac{e^\theta - 1}{\theta^2}$$

Достаточные статистики

3.5

Опр Пусть $X = (X_1, \dots, X_n)$ — выборка из $P \in \mathcal{P}$

Статистика $S(X)$ наз-ся достаточной для семейства $\mathcal{P}$, если условное распр. $P_\theta(X \in B \mid S(X))$ не зависит от θ $\forall B$

Смысл: вся информация о θ, содержащаяся в выборке содержится в достаточной статистике

Тривиальный пример: $S(X) = (X_1, \dots, X_n)$

Следствие: Если данные поступают последовательно, то достаточно хранить только $S(X)$

Важным является случай, когда размерность достаточной статистики меньше размера выборки. $|S(x)| \ll n$

[на самом деле такое $S(x)$ не всегда сущ.
например для распр. Коши]

Пример $X_1 \ldots X_n \sim \text{Bern}(\theta)$

Какая информация есть в выборке?

Кол-во успехов $S(x) = \sum\limits_{i=1}^{n} X_i$

Порядок успех-неудач (какой-то убийцный порядок без количеств)

$S(x)$ — достаточная

$$P_\theta(X = \bar{x} \mid S(X) = s) = \frac{P_\theta(X_1 = x_1 \ldots X_n = x_n, \sum X_i = s)}{P_\theta(\sum X_i = s)} = \left[\begin{array}{l}\text{если } \sum x_i \neq s, \\ \text{то } 0\end{array}\right]$$

$$= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\, I\{\sum x_i = s\}}{\theta^s (1-\theta)^{n-s}\, C_n^s} = \frac{I\{\sum x_i = s\}}{C_n^s} \leftarrow \text{не зависит от } \theta$$

Теор. (Критерий факторизации Неймана - Фишера)

Пусть $X = (X_1, \dots, X_n)$ — выборка из распр $P \in \mathcal{P}$ - домин. семейство

Тогда $S(X)$ — дос. стат. $\Leftrightarrow$ справедлива факторизация

$$p_\theta(X) = \psi(S(X), \theta) \cdot \underbrace{h(x)}_{\text{не зависит от } \theta}$$

Пример $X_1 \dots X_n \sim \Gamma(\alpha, \beta) \qquad \theta = (\alpha, \beta)$

$$p_\theta(x) = \frac{\alpha^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha x}$$

$$p_\theta(x_1 \dots x_n) = \frac{\alpha^{\beta n}}{\Gamma(\beta)^n} \left(\prod_{i=1}^n x_i\right)^{\beta-1} e^{-\alpha \sum x_i} = \psi(\underbrace{S(x)}_{(\sum x_i,\ \prod x_i)}, \underbrace{\theta}_{(\alpha,\beta)}) \cdot h(x)$$

$$\psi(x, y, \alpha, \beta) = \frac{\alpha^{\beta n}}{\Gamma(\beta)^n} y^{\beta-1} e^{-\alpha x}$$

$$h(x) = 1$$

$$S(x) = \left(\sum x_i, \prod x_i\right)$$

$\Downarrow$

достаточные статистики

$$S(x) = \left(\sum x_i, \sum \ln x_i\right)$$