Matematika III – 13. týden
Lineární modely

Jan Slovák

Masarykova univerzita
Fakulta informatiky

9.12. - 13.12. 2014

Obsah přednášky

1. Literatura
2. Lineární modely
3. Vzpomníky na lineární algebru
4. Vícerozměrné $N_m(\mu, V)$
5. Hlavní věta
6. Použití

Plán přednášky

1. Literatura
2. Lineární modely
3. Vzpomníky na lineární algebru
4. Vícerozměrné $N_m(\mu, V)$
5. Hlavní věta
6. Použití

Kde je dobré číst?

- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- J. Slovák, M. Panák, M. Bulant, Matematika drsně a svižně, Muni Press, Brno 2013, v+773 s., elektronická edice `www.math.muni.cz/Matematika_drsne_svizne`
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, Teorie pravděpodobnosti a matematická statistika (sbírka příkladů), Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, Základní statistické metody, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.
- Riley, K.F., Hobson, M.P., Bence, S.J. Mathematical Methods for Physics and Engineering, second edition, Cambridge University Press, Cambridge 2004, ISBN 0 521 89067 5, xxiii + 1232 pp.

Plán přednášky

1. Literatura
2. Lineární modely
3. Vzpomníky na lineární algebru
4. Vícerozměrné $N_m(\mu, V)$
5. Hlavní věta
6. Použití

Uvažujme náhodný vektor $Y = (Y_1, \ldots, Y_n)^T$ a předpokládejme, že platí

$$Y = X \cdot \beta + \sigma Z,$$

kde $X = (x_{ij})$ je konstantní matice reálných čísel s n řádky a $k < n$ sloupci a hodností k, β je neznámý konstantní vektor k parametrů modelu, Z je náhodný vektor, jehož n komponent má rozdělení $\mathrm{N}(0, 1)$, a $\sigma > 0$ je neznámý kladný parametr modelu. Hovoříme o **lineárním modelu** s úplnou hodností.

Uvažujme náhodný vektor $Y = (Y_1, \ldots, Y_n)^T$ a předpokládejme, že platí

$$Y = X \cdot \beta + \sigma Z,$$

kde $X = (x_{ij})$ je konstantní matice reálných čísel s n řádky a $k < n$ sloupci a hodností k, β je neznámý konstantní vektor k parametrů modelu, Z je náhodný vektor, jehož n komponent má rozdělení $N(0, 1)$, a $\sigma > 0$ je neznámý kladný parametr modelu. Hovoříme o **lineárním modelu** s úplnou hodností.

V praktických problémech zpravidla známe veličiny x_{ij} a snažíme se odhadnout nebo predikovat hodnotu Y.

Chceme přitom mít jasno o pravděpodobnostních charakteristikách těchto odhadů.

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i–tého studenta v j–tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i–tého studenta v j–tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).
K odhadu vektoru β se často používá **metoda nejmenších čtverců**. To znamená, že chceme najít odhad $b \in \mathbb{R}^k$ pro vektor β tak, aby vektor $\hat{Y} = Xb$ minimalizoval druhou mocninu délky vektoru $Y - X\beta$.

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i–tého studenta v j–tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).
K odhadu vektoru β se často používá **metoda nejmenších čtverců**. To znamená, že chceme najít odhad $b \in \mathbb{R}^k$ pro vektor β tak, aby vektor $\hat{Y} = Xb$ minimalizoval druhou mocninu délky vektoru $Y - X\beta$.
To je ale jednoduchá úloha lineární algebry a víme, že jde o nalezení kolmého průmětu vektoru Y do podprostoru $\langle X \rangle \subset \mathbb{R}^n$ generovaném sloupci matice X.

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2.$$

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2.$$

Velikost $\|Y - \hat{Y}\|^2$ nazýváme **reziduální součet čtverců**, zpravidla se značí **RSS**. Definujeme také **reziduální rozptyl** jako

$$S^2 = \frac{\|Y - Xb\|^2}{n - k}.$$

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2.$$

Velikost $\|Y - \hat{Y}\|^2$ nazýváme **reziduální součet čtverců**, zpravidla se značí **RSS**. Definujeme také **reziduální rozptyl** jako

$$S^2 = \frac{\|Y - Xb\|^2}{n - k}.$$

Víme, že $\hat{Y} = Xb$ a že, díky našemu přepokladu o maximální hodnosti X, je matice $X^T X$ invertibilní. Můžeme proto rovnou spočíst $b = (X^T X)^{-1} X^T \hat{Y}$.

Plán přednášky

1 Literatura

2 Lineární modely

3 Vzpomníky na lineární algebru

4 Vícerozměrné $N_m(\mu, V)$

5 Hlavní věta

6 Použití

Theorem

Nechť A je libovolná matice typu m/n nad reálnými nebo komplexními skaláry. Pak existují čtvercové unitární matice U a V dimenzí m a n, a reálná diagonální matice s nezápornými prvky D dimenze r, $r \leq \min\{m, n\}$, takové, že

$$A = USV^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

a r je hodnost matice AA^*. Přitom je S určena jednoznačně až na pořadí prvků a prvky diagonální matice D jsou druhé odmocniny vlastních čísel d_i matice AA^*. Pokud je A reálná matice, pak i matice U a V jsou ortogonální.

Definition

Nechť A je reálná matice typu m/n a nechť

$$A = USV^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

je její singulární rozklad (zejména D je invertibilní). Matici

$$A^\dagger := VS'U^*, \quad S' = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

nazýváme **pseudoinverzní matice** k matici A.

Jak ukazuje následující věta, je pseudoinverze důležité zobecnění pojmu inverzní matice, včetně přímočarých aplikací.

Theorem

Nechť A je reálná nebo komplexní matice typu m/n. Pak pro její pseudoinverzní matici platí:

1. *Je-li A invertibilní (zejména tedy čtvercová), pak*
$$A^\dagger = A^{-1}.$$
2. *Pro pseudoinverzi $A^\dagger$ platí, že $A^\dagger A$ i $AA^\dagger$ jsou hermiteovské (v reálném případě symetrické) a*
$$AA^\dagger A = A, \quad A^\dagger AA^\dagger = A^\dagger.$$
3. *Pseudoinverzní matice $A^\dagger$ je čtyřmi vlastnosti z předchozího bodu určena jednoznačně. Pokud tedy nějaká matice B typu $n \times m$ splňuje, že BA i AB jsou hermiteovské, $ABA = A$ a $BAB = B$, pak $B = A^\dagger$.*

Theorem (Pokračování)

1. *Je-li A matice systému lineárních rovnic $Ax = b$, s pravou stranou $b \in \mathbb{K}^m$, pak vektor $y = A^\dagger b \in \mathbb{K}^n$ minimalizuje velikost $\|Ax - b\|$ pro všechny vektory $x \in \mathbb{K}^n$.*
2. *Systém lineárních rovnic $Ax = b$ s $b \in \mathbb{K}^m$ je řešitelný, právě když platí $AA^\dagger b = b$. V tomto případě jsou všechna řešení dána výrazem*

$$x = A^\dagger b + (E - A^\dagger A)u,$$

kde $u \in \mathbb{K}^n$ je libovolné.

Z bodu (4) předchozí věty plyne, že matice $AA^\dagger$ je maticí kolmé projekce z vektorového prostoru $\mathbb{R}^n$, kde n je počet řádků matice A na podprostor generovaný sloupci matice A (tato interpretace má samozřejmě smysl pouze pro matice mající více řádků než sloupců). Dále pro matice A, jejichž sloupce tvoří nezávislé vektory, má smysl výraz $(A^T A)^{-1} A^T$ a není těžké ověřit, že tato matice splňuje všechny vlastnosti z (1) a (2) z předchozí věty, jedná se tedy o pseudoinverzi k matici A.

Z předchozí věty vyplývají také následující vlastnosti pseudoinverze:

- Pro všechny matice A platí $(A^\dagger)^\dagger = A$,
- pokud má matice A, typu $m \times n$, plnou řádkovou hodnost m, pak $A^\dagger = A^*(AA^T)^{-1}$,
- pokud má matice A, typu $m \times n$, plnou sloupcovou hodnost n, pak $A^\dagger = (A^TA)^{-1}A^*$.

Plán přednášky

1. Literatura
2. Lineární modely
3. Vzpomníky na lineární algebru
4. Vícerozměrné $N_m(\mu, V)$
5. Hlavní věta
6. Použití

Jestliže má náhodný vektor $Z = (Z_1, \ldots, Z_n)$ nezávislé komponenty $Z_i \sim \mathrm{N}(0, 1)$, je jeho varianční matice jednotkovou maticí, tj. $\mathrm{var}\, Z = \mathbb{I}_n$.
Uvažme vektor $U = a + BZ$, kde a je libovolný konstantní vektor v $\mathbb{R}^m$ a B je konstantní matice typu (m, n).
Víme $\mathrm{E}\, U = a$ a $\mathrm{var}\, U = V = BB^T$ (protože varianční matice Z je identická). Je tedy tato varianční matice vždy pozitivně semidefinitní.
Říkáme, že náhodný vektor U má **mnohoměrné normální rozdělení** $\mathrm{N}_m(a, V)$.

Pro libovolné mnohoměrné normální rozdělení $\mathrm{N}_m(a, V)$ znovu vezmeme afinní transformaci

$$W = c + DU$$

s vektorem konstant $c \in \mathbb{R}^k$ a libovolnou konstantní maticí typu (k, m). Přímým výpočtem

$$W = c + D(a + BZ) = (c + Da) + (DB)Z,$$

což je samořejmě náhodný vektor $W \sim \mathrm{N}_k(c + Da, DB^T BD^T)$. Chová se tedy kovarianční matice mnohoměrného normálního rozdělení při afinních transformacích jako kvadratická forma.

Dokázali jsme, že jakákoliv linární kombinace komponent složek náhodného vektoru s mnohoměrným normálním rozdělením je náhodná veličina s normálním rozdělením. Stejně je každý vektor vzniklý výběrem jen některých komponent vektoru U opět náhodným vektorem s mnohoměrným normálním rozdělením.

Plán přednášky

1. Literatura
2. Lineární modely
3. Vzpomníky na lineární algebru
4. Vícerozměrné $N_m(\mu, V)$
5. Hlavní věta
6. Použití

Theorem

V lineárním modelu $Y = X\beta + \sigma Z$ platí pro vhodné matice P a R:
(1) Pro odhad $\hat{Y}$ platí

$$\hat{Y} = X\beta + \sigma PP^T Z, \quad \hat{Y} \sim \mathsf{N}(X\beta, \sigma^2 PP^T).$$

(2) Reziduální součet čtverců RSS a normovaný čtverec velikosti rezidua mají rozdělení:

$$Y - \hat{Y} \sim \mathsf{N}(0, \sigma^2 RR^T), \quad \|Y - Y\|^2/\sigma^2 \sim \chi^2_{n-k}\,.$$

(3) Náhodná veličina $b = \beta + \sigma(P^T X)^{-1} P^T Z$ má rozdělení

$$b \sim \mathsf{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

(4) Pro reziduální rozptyl platí $(n-k)S^2/\sigma^2 \sim \chi^2_{n-k}$.
(5) Střední hodnota reziduálního rozptylu je $\mathsf{E}\, S^2 = \sigma^2$.
(6) Veličiny b a S^2 jsou nezávislé.

Plán přednášky

1. Literatura
2. Lineární modely
3. Vzpomníky na lineární algebru
4. Vícerozměrné $N_m(\mu, V)$
5. Hlavní věta
6. Použití

Úplně nejjednodušší je to v případě jediného výběru, kdy testujeme, zda jediný parametr β je roven dané hodnotě β_0.
Volíme matici X s jediným sloupcem plným jedniček. Výraz

$$Y = X\beta + \sigma Z$$

komponenty v Y jsou nezávislé veličiny $Y_i \sim N\beta, \sigma^2$, jde o náhodný výběr rozsahu n z normálního rozdělení.
Obecná věta dává odhad

$$b = (X^T X)^{-1} X^T Y = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$$

$$S^2 = \frac{1}{n-1} \|Y - X\bar{Y}\|^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

což jsou právě výběrový průměr a rozptyl, se kterými jsme již počítali.

Zajímá nás přitom statistika

$$T = \frac{\bar{Y} - \beta_0}{S}\sqrt{n}$$

Testování hypotézy $\beta = \beta_0$ se nazývá **jednovýběrový t-test**. Na hladině α hypotézu zamítáme, když je $|T| \geq t_{n-1}(\alpha)$.

párový t-test

Je vhodný na případy, kdy testujeme dvojice náhodných vektorů $W_1 = (W_{i1})$ a $W_2 = (W_{i2})$, o rozdílech jejichž komponent $Y_i = W_{i1} - W_{i2}$ víme, že mají rozdělení $\mathrm{N}(\beta, \sigma^2)$. Potřebujeme navíc, aby byly veličiny Y_i nezávislé (což neříká, že musí být nezávislé jednotlivé dvojice W_{i1} a W_{i2}!). Můžeme si představit třeba hodnocení dvou různých vyučujících tímž studentem. Testujeme hypotézu, že pro všechna i je $\mathrm{E}\, W_{i1} = \mathrm{E}\, W_{i2}$. Používáme statistiku

$$T = \frac{\bar{W}_1 - \bar{W}_2}{S}\sqrt{n}.$$