

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Riadkovo-stĺpcové návrhy štatistických experimentov

BAKALÁRSKA PRÁCA

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Riadkovo-stĺpcové návrhy štatistických experimentov

BAKALÁRSKA PRÁCA

Študijný program: Matematika
Študijný odbor: 1114 Matematika
Školiace pracovisko: Katedra aplikovanej matematiky a štatistiky
Vedúci práce: doc. Mgr. Radoslav Harman, PhD.



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Róbert Druska
Študijný program: matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Riadkovo-stĺpcové návrhy štatistických experimentov
Row-column designs of statistical experiments

Anotácia: Prvým cieľom je analyzovať vlastnosti odhadov parametrov regresného modelu pre takzvaný riadkovo-stĺpcový experimentálny návrh. Druhým cieľom je navrhnúť algoritmus na výpočet optimálneho návrhov tohto typu, v závislosti na požiadavkách experimentátora.

Vedúci: doc. Mgr. Radoslav Harman, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.
Dátum zadania: 15.10.2019

Dátum schválenia: 18.10.2019

prof. RNDr. Ján Filo, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Pod'akovanie TODO

Abstrakt

TODO

Kľúčové slová: lineárny regresný model

Abstract

TODO

Keywords: linear regression

Obsah

Úvod	8
1 Lineárny regresný model	9
1.1 Metóda najmenších štvorcov	9
1.2 Metodika skúmania	11
2 Môj model	12
Záver	15
Zoznam použitej literatúry	16
Príloha A	17

Úvod

TODO

1 Lineárny regresný model

Majme n nameraných štatistických jednotiek tvaru $\{y, x_1, \dots, x_p\}$, ktoré sme dostali ako výsledok experimentu. Lineárny regresný model predpokladá, že medzi jednotlivými prvkami y, x_1, \dots, x_p je lineárny vzťah. Motiváciou za lineárnym regresným modelom je spravidla aproximovať tento lineárny vzťah.

Aproximácia lineárneho vzťahu nám v praxi ponúkne mechanizmus, ktorým možno predikovať neznámu hodnotu y na základe známych hodnôt y, x_1, \dots, x_p , čo v reálnom živote predstavuje často sa vyskytujúci problém.

Označme teda daný lineárny vzťah medzi zložkami nameranej štatistickej jednotky:

$$y_i = b_0 + b_1 x_{i1} + \dots + b_p x_{ip} + e_i = b^T x_i + e_i$$

kde $\{y_i, x_i\}$ je i -ta nameraná jednotka, b je vektor lineárneho vzťahu a e_i je chyba merania.

Keď lineárne vzťahy pre každú z n nameraných jednotiek zapíšeme maticovo, dostaneme vzťah

$$y = Xb + e \tag{1}$$

kde $y = (y_1, y_2, \dots, y_n)^T$, $e = (e_1, e_2, \dots, e_n)^T$ a

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

je matica tvaru $n \times p$. V praxi je b neznámy vektor, ktorý sa snažíme odhadnúť.

Na spočítanie odhadu b sa používajú rôzne metódy, najčastejšie napr. metóda najmenších štvorcov alebo metóda maximálnej vierohodnosti.

1.1 Metóda najmenších štvorcov

Metódou najmenších štvorcov vypočítame odhad \hat{b} parametra b nasledovne:

$$\hat{b} = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} (y - Xb)^T C (y - Xb) = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \|y - Xb\|_{C^{-1}}^2$$

kde C je nejaká kladne definitná matica. Ak $C = I$, potom minimalizujeme výraz $\|y - Xb\|_I^2 = \|y - Xb\|^2 = \sum_{i=1}^n (y_i - X_i \cdot b)^2$, kde X_i značí i -ty riadok matice X . V našej práci budeme predpokladať homogenitu chýb, čo v praxi znamená, že skutočne budeme môcť dosadiť $C = I$. Preto maticu C v ďalšom opise teórie spomínať nebudeme.

Geometricky metódu najmenších štvorcov možno interpretovať ako projekciu vektora y na stĺpcový priestor matice X . Hľadáme teda taký vektor \hat{b} , pre ktorý platí $X\hat{b} = Py$, kde P je matica ortogonálnej projekcie na stĺpcový priestor X . Z teórie lineárnej algebry vieme, že $P = X(X^T X)^- X^T$, kde znamienko $-$ označuje g -inverziu. (g -inverziou matice A je taká matica A^- , pre ktorú platí $AA^-A = A$).

Odhad \hat{b} parametra b je teda riešením rovnice

$$Xb = X(X^T X)^- X^T y \quad (2)$$

Toto riešenie spočítame ako:

$$\hat{b} = (X^T X)^- X^T y$$

kde použitá g -inverzia je ľubovoľná.

Z uvedeného vyplýva, že v prípade regulárnosti X je odhad \hat{b} jednoznačný. V našej práci budeme skúmať matice (modely) X , ktoré nie sú regulárne, takže jednoznačný odhad \hat{b} nebudeme schopní nájsť (čo v konečnom dôsledku ani nie je naším záujmom). Budeme odhadovať lineárnu funkciu zložiek vektora b , konkrétne $h^T b = h_1 b_1 + \dots + h_p b_p$, ktorá býva odhadnuteľná aj v prípade singularity X , ak vektor h spĺňa určité podmienky. Je niekoľko ekvivalentných podmienok, ktoré stačia na to, aby $h^T b$ bolo odhadnuteľné. Z nich spomenieme jednu v nasledovnej vete, ktorú použijeme neskôr v našej práci.

Veta 1.1. $h^T b$ je odhadnuteľné, ak platí nasledovné ekvivalentné podmienky:

1. pre ľubovoľné riešenia b^* a b^{**} rovnice (2) platí $h^T b^* = h^T b^{**}$
2. $h \in \mathcal{M}(X^T)$
3. $h \in \mathcal{M}(X^T X)$,

kde \mathcal{M} označuje stĺpcový priestor matice.

Ak h patrí do riadkového priestoru matice X , potom existuje také u , že $h = F^T u$. Potom pre jednoznačný odhad $h^T \hat{b}$ vektora $h^T b$ platí:

$$h^T \hat{b} = u^T X \hat{b} = u^T P y = u^T X (X^T X)^- X^T y$$

Výsledok predchádzajúcej vety je dôležitý pre našu prácu, pretože nebudeme skúmať odhady b , ale odhady niektorých lineárnych kombinácií zložiek vektora b , konkrétne napr. rozdiely medzi parametrami.

Odhad \hat{b} parametra b , ako aj odhad $h^T \hat{b}$ parametra $h^T b$, sú lineárne nevychýlené odhady, ktorým prislúcha disperzia (TODO: popremýšľaj, či treba bližšie opísať lineárny nevychýlený odhad). Neskôr v našej práci budeme hľadať také modely X , pri ktorých je disperzia odhadov b či $h^T b$ najmenšia možná, čo nám dá najlepší lineárny nevychýlený odhad. Ak nami navrhované modely X budú opisovať ten istý experiment, ten model X , pre ktorý disperzia odhadu $h^T b$ bude najmenšia, bude svojím spôsobom optimálny.

K nájdeniu optimálneho modelu X nám poslúži Gaussova-Markovova veta, ktorá určuje minimálnu možnú disperziu odhadu $h^T b$.

Veta 1.2. (*Gaussova-Markovova*) *Nech h je z riadkového priestoru X . Potom minimálna možná disperzia lineárneho nevychýleného odhadu $h^T b$ je*

$$m = \text{Var}[h^T b] = h^T M^- h$$

kde $M = X^T X$ je informačná matica parametra b a M^- je jej ľubovoľná g -inverzia.

1.2 Metodika skúmania

Cieľom nášho skúmania bude nájsť ideálnu maticu vstupu do experimentu (konkrétny problém popíšeme v nasledujúcej kapitole). Pokúsime sa vysloviť závery pre všeobecný prípad matice $m \times n$, avšak pre istý prvotný náhľad do problematiky problém popíšeme a preskúmame na menších rozmeroch, konkrétne do rozmeru, do ktorého nám to umožní výpočtová technika. Pracovať budeme s programovacím jazykom *python*.

Od „preskúmania situácie“ do istého rozmeru si sľubujeme dôležité náhľady, ktoré nám umožnia vysloviť závery a hypotézy pre všeobecný prípad $m \times n$. Tie sa následne pokúsime dokázať matematickými metódami.

2 Môj model

Cieľom tejto práce je skúmať experiment reprezentovaný binárnou maticou. Možná reprezentácia nášho modelu pri matici rozmeru $m \times n$ je takáto:

Uvažujeme model s dvomi kvalitatívnymi faktormi A a B , pričom faktor A má m úrovní a faktor B n úrovní. Pre každú kombináciu faktorov experimentátor zvolí jedno z dvojice ošetrení (angl. treatment). Budeme uvažovať len aditívny model bez interakcií.

(Harmanov text == TODO: prepísať) Napríklad faktor A môže označovať dobrovoľníka (t.j. máme m dobrovoľníkov) a faktor B účinnú látku, ktorú nazveme liečivo (t.j. máme n liečiv). Ošetrovanie 0 reprezentuje podanie liečiva orálnym spôsobom a ošetrovanie 1 vnútrožilne). Účinok, ktorý po čase nameriame na dobrovoľníkovi, závisí aditívne od efektu samotného dobrovoľníka (jeho predispozícií), efektu liečiva a efektu ošetrovania. Všetky tieto efekty uvažujeme ako neznáme parametre modelu.

Návrh teda možno reprezentovať binárnou maticou (alebo bipartitným grafom, čo opíšeme neskôršie v našej práci).

Cieľom experimentu je odhadnúť rozdiel medzi dvoma efektmi (treatmentami). Cieľom hľadania optimálneho modelu je nájsť taký návrh, pri ktorom rozptyl Gauss-Markovovho odhadu rozdielu medzi efektmi dvojice ošetrení (t.j. odhadu kontrastu ošetrení) bude čo najmenší. (Použijeme pri tom Gaussovu-Markovovu vetu z predošlej kapitoly.)

Pre maticu typu $m \times n$ dostaneme teda mn vzťahov tvaru:

$$y_{ij} = a_i + b_j + t + e_{ij}$$

$i = 1, \dots, m$; $j = 1, \dots, n$; t je t_0 alebo t_1 podľa hodnoty na ij -tej pozícii matice a e_{ij} je chyba. Budeme odhadovať rozdiel medzi t_0 a t_1 .

Z návrhu modelu v podobe binárnej matice tvaru $m \times n$ vytvoríme lineárny regresný model s maticou tvaru $mn \times (m+n+2)$, pretože máme mn meraní závislých od $m+n+2$ premenných.

Príklad:

Maticu návrhu experimentu

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad (3)$$

prepíšeme ako (TODO: popíš tými strapatými zátvorkami čo znamená čo):

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad (4)$$

Označme vo všeobecnosti maticu modelu (3) B a maticu lineárneho zobrazenia (4) X . Predpokladáme lineárne vzťahy, preto dostávame lineárnu regresiu

$$y = Xb + e$$

kde y je vektor dát nameraných v experimente, b je vektor neznámych koeficientov a $e = (e_{11}, e_{12}, e_{13}, \dots, e_{m1}, \dots, e_{mn})^T$ je vektor chýb merania. Uvedomme si, že v našej práci nebudeme pracovať s konkrétnym vektorom y ; snažíme sa totiž navrhnúť maticu modelu B (a z nej vyplývajúcu maticu X) ešte pred tým, než experiment vykonáme.

Naším cieľom teraz bude odhadnúť rozdiel medzi efektmi t_0 a t_1 , resp. zistiť, aká môže byť disperzia tohto odhadu pri danej matici návrhu B . Modely, pre ktoré bude možná disperzia najmenšia, budeme považovať za optimálne.

V danej lineárnej regresii $y = Xb + e$ teda nebudeme odhadovať celý vektor b , ale len jeho lineárnu kombináciu $h^T b$, kde h je vektor rozmeru $m + n + 2$, ktorý má na $m + n$ miestach 0 a na zvyšných dvoch miestach +1 a -1, ktoré zodpovedajú efektom t_0 a t_1 . Z algoritmu, ktorým sme z matice návrhu B vytvorili maticu lineárnej regresie X , vyplýva, že vektor h má tvar $(0, 0, 0, \dots, +1, -1)^T$.

Na zistenie, či $h^T b$ je odhadnuteľné, použijeme vetu (1.1), a na následné nájdenie minimálnej možnej disperzie použijeme Gaussovu-Markovovu vetu (1.2).

Záver

TODO

Zoznam použitej literatúry

- [1] Pázman, A., Lacko, V.: *Prednášky z regresných modelov*, Vydavateľstvo UK, Bratislava, 2012, 2015

Príloha A