UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Riadkovo-stĺpcové návrhy štatistických experimentov

BAKALÁRSKA PRÁCA

2020 Róbert Druska

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.

Bratislava 2020 **Róbert Druska**

34792200

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 parameterov 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

Poďakovanie TODO

Abstrakt

TODO

Kľúčové slová: lineárny regresný model

Abstract

TODO

Keywords: linear regression

Obsah

Úvod

TODO

1 Lineárny regresný model

Majme n nameraných štatistických jednotiek tvaru $\{y, x_1, \ldots, x_p\}$, ktoré sme dostali ako výsledok experimentu. Lineárny regresný model predpokladá, že medzi jednotlivými prvkami $y, x_1, ...x_p$ 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 $y, x_1, ...x_p$, čo v reálnom živote predstavuje často sa vyskytujúci problém.

Označme teda daný lineárny vzťah medzi zložkami nameranej štatistickej jednotky:

$$y_i = b_0 + b_1 x_{i_1} + \ldots + b_p x_{i_p} + e_i = b^T x_i + e_i$$

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 $y = (y_1, y_2, \ldots, y_n)^T$, $e = (e_1, e_2, \ldots, e_3)^T$ a

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

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:

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

kde C je nejaká kladne definitná matica. Ak $C = I$, potom minimalizujeme výraz $||y-Xb||^2_I = ||y-Xb||^2 = \sum_{i=1}^{n} (y_i - X_{i\cdot} b)^2$, kde $X_{i\cdot}$ 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 $P = X(X^TX)^-X^T$, 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

$$Xb = X(X^TX)^-X^Ty \tag{2}$$

Toto riešenie spočítame ako:

$$\hat{b} = (X^TX)^-X^Ty$$

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 $h^Tb = h_1b_1 + \ldots + h_pb_p$, ktorá býva odhadnuteľná aj v prípade singularity X, ak vektor h spĺňa určité podmienky. Je niekoľko ekvivalentných podmienok, ktoré stačia na to, aby h^Tb bolo odhadnuteľné. Z nich spomenieme jednu v nasledovnej vete, ktorú použijeme neskôr v našej práci.

Veta 1.1. *h^Tb je odhadnuteľné, ak plati nasledovné ekvivaltentné podmienky:*

1. *pre ľubovoľné riešenia b^* a b^{**} rovnice (2) platí $h^Tb^* = h^Tb^{**}$*

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:

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

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*

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

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

2 ...

Záver

TODO

Zoznam použitej literatúry

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

Príloha A